INTERNATIONAL DOCTORAL SCHOOL OF THE USC

Manuel
Couto Pintos

PhD Thesis

Online content mining to improve psychological profiling of individuals

Santiago de Compostela, 2026

Doctoral Programme in Information Technology Research

ESCOLA DE DOUTORAMENTO
INTERNACIONAL DA USC

DOCTORAL THESIS

ONLINE CONTENT MINING TO IMPROVE PSYCHOLOGICAL PROFILING OF INDIVIDUALS

Author

Manuel Couto Pintos

Supervisor/s: David E. Losada Carril, Javier Parapar López

Tutor: David E. Losada Carril

PHD PROGRAMME IN INFORMATION TECHNOLOGY RESEARCH

SANTIAGO DE COMPOSTELA - 2026

The PhD candidate declares that they have no conflict of interest in relation to the doctoral thesis.

Agradecementos

Gustaríame agradecer á miña familia e amigos o seu apoio durante estes anos. Aos compañeiros do CiTIUS e do CITIC cos que compartín tempo durante a miña etapa como investigador. E aos meus directores de tese, David e Javier, polo seu apoio e orientación durante a realización desta tese.

Este traballo recibiu financiamento de MICIU/AEI/10.13039/501100011033 (PID2022-137061OB-C22, cofinanciado polo FEDER) e da Xunta de Galicia-Consellería de Cultura, Educación, Formación Profesional e Universidades (ED431G 2023/04, ED431C 2022/19, cofinanciados polo FEDER).

2026

Contents

Resumo

Os trastornos de saúde mental constitúen un dos maiores desafíos de saúde pública do século XXI. Segundo a Organización Mundial da Saúde, máis de 280 millóns de persoas en todo o mundo padecen depresión, e a prevalencia de condicións como a ansiedade, os trastornos alimentarios, as autolesións e o xogo patolóxico segue en aumento. A detección temperá e a intervención oportuna son amplamente recoñecidas como factores críticos para mellorar os resultados clínicos; con todo, as vías diagnósticas tradicionais seguen limitadas pola escaseza de recursos, o estigma social e a dificultade inherente de identificar o malestar psicolóxico antes de que derive nunha crise clínica. A carga económica, estimada en máis dun billón de dólares anuais en produtividade perdida a nivel mundial, súmase ao custo humano e subliña a urxencia de desenvolver ferramentas escalables e complementarias que poidan apoiar as vías clínicas tradicionais sen substituílas.

Ao mesmo tempo, as plataformas de redes sociais convertéronse en espazos nos que os usuarios revelan experiencias persoais cunha franqueza notable, moitas veces ao abeiro do anonimato que estas plataformas proporcionan. Os usuarios discuten con frecuencia as súas dificultades con depresión, trastornos alimentarios e outras condicións de saúde mental de maneiras que serían difíciles de captar mediante instrumentos clínicos convencionais. A investigación en psicolingüística estableceu que os patróns lingüísticos correlacionan con estados psicolóxicos, e as redes sociais proporcionan unha ventá sen precedentes a estes patróns a gran escala. A converxencia destes dous fenómenos —a crise de saúde mental e a proliferación da autoexpresión dixital— deu lugar a unha área de investigación en rápido crecemento na intersección do Procesamento da Linguaxe Natural, a Recuperación de Información e a Psicoloxía Computacional. Iniciativas como as tarefas compartidas eRisk no marco de CLEF e os talleres CLPsych catalizaron os esforzos da comunidade ao proporcionar conxuntos de datos estandarizados, protocolos de avaliación e bancos de proba competitivos para a detección temperá de

riscos de saúde mental a partir de texto de redes sociais. Esta oportunidade, sen embargo, leva consigo responsabilidades éticas e metodolóxicas substanciais, particularmente arredor da preservación da privacidade, a evitación de enfoques estigmatizantes e o recoñecemento de que os sinais computacionais deben informar —e non substituír— o xuízo clínico profesional. Nembargantes, o campo segue enfrontándose a desafíos fundamentais en representación textual, descubrimento de patróns, avaliación de modelos e reproducibilidade experimental.

Esta tese, presentada como un compendio de cinco publicacións, desenvolve e avalía métodos computacionais para a análise e detección temperá de condicións de saúde mental a partir de texto de redes sociais. As contribucións abarcan avances en agrupamento de textos, modelado de temas, avaliación da coherencia, word embeddings temporais e experimentación reproducible. O traballo estrutúrase arredor de cinco hipóteses de investigación, cada unha abordada por unha publicación específica. A fonte de datos principal ao longo da tese é a colección de datos experimentais publicados pola iniciativa eRisk, que comprende publicacións e comentarios recollidos de Reddit, onde os usuarios foron identificados como portadores de signos de condicións específicas de saúde mental ou como pertencentes a un grupo de control. Catro condicións de saúde mental están cubertas: depresión (edicións eRisk 2017, 2018 e 2022), anorexia (2018 e 2019), autolesións (2020 e 2021) e trastornos de xogo (2022 e 2023). Estes conxuntos de datos presentan características que os fan particularmente desafiantes para a análise computacional: a distribución de clases está fortemente desequilibrada, as publicacións están escritas en inglés informal cunha variación significativa en lonxitude e estilo, e os historiais de publicación dos usuarios abarcan períodos estendidos que introducen variabilidade e factores estacionais. De media, cada usuario achega centos de publicacións distribuídas ao longo de meses ou anos, proporcionando un trazo lonxitudinal do comportamento lingüístico que sería difícil de obter por calquera outro medio. As representacións textuais empregadas ao longo da tese van desde métodos clásicos dispersos (bolsa de palabras, TF-IDF) ata embeddings densos estáticos (Word2Vec, GloVe, FastText) e embeddings contextuais de modelos transformadores (BERT, RoBERTa, ALBERT, MPNET). A aplicación de algoritmos de agrupamento e modelado de temas a este tipo de datos enfronta o problema da maldición da dimensionalidade, o que levou á adopción xeneralizada de arquitecturas de pipeline que descompoñen a tarefa en etapas: representación textual, redución de dimensionalidade (mediante métodos como UMAP) e agrupamento ou modelado sobre as representacións reducidas.

A primeira publicación leva a cabo unha comparación sistemática de algoritmos de agrupamento –baseados en particións (K-Means), baseados en modelos (Modelos de Mesturas Gau-

sianas, GMM) e baseados en densidade (DBSCAN)– para o descubrimento de coñecemento en datos de Reddit relacionados coa depresión, empregando o conxunto de datos de eRisk 2017. O estudo emprega múltiples representacións textuais, incluíndo vectores de frecuencia de termos (TF), TF-IDF e embeddings contextuais de modelos transformadores (BERT, RoBERTa, GPT-2). Aplicouse un preprocesamento estándar que inclúe tokenización, filtrado de palabras baleiras e selección por lonxitude antes da representación, e os hiperparámetros de cada algoritmo de agrupamento exploráronse en intervalos razoables para garantir que as diferenzas observadas reflectisen os métodos subxacentes e non escollas incidentais de axuste. Os resultados revelaron que a mellor configuración –GMM con embeddings de RoBERTa– xenera agrupamentos correspondentes a relacións persoais, videoxogos e discusións sobre substancias, mostrando o agrupamento de relacións unha maior proporción de usuarios deprimidos. Non obstante, o agrupamento non separou de forma efectiva as categorías diagnósticas (ARI = 0,180, NMI = 0,200), confirmando que a depresión non se manifesta como unha clase xeometricamente separable en ningún dos espazos de embeddings avaliados.

O achado máis significativo desta publicación foi de natureza metodolóxica: as métricas de validación interna (Coeficiente de Silueta, Calinski-Harabasz, Davies-Bouldin) non son comparables entre espazos de representación con propiedades xeométricas diferentes. Nos nosos experimentos, os vectores TF –cuxa magnitude crece coa lonxitude do documento– maximizaron consistentemente as métricas internas mentres que obtiveron o peor rendemento na validación externa. Confiar inxenuamente nas métricas internas tería levado á selección do peor método de representación, o que ilustra un modo de fallo que probablemente afecte a calquera estudo que compare agrupamentos entre espazos de embeddings heteroxéneos. As implicacións esténdense moito máis alá da análise de saúde mental, afectando a calquera aplicación na que se avalíe a calidade do agrupamento entre representacións heteroxéneas, incluíndo sistemas de recomendación, organización documental e minería de texto biomédico.

As limitacións do enfoque de agrupamento motivaron unha evolución natural cara ao modelado de temas. Ambas técnicas comparten un eixo computacional común –representación mediante embeddings, redución de dimensionalidade e agrupamento– pero o modelado de temas engade unha capa de descrición que extrae termos representativos para cada grupo, transformando agrupamentos opacos en temas interpretables. A segunda publicación explora esta dirección, comparando LDA, BERTopic e TopClus nos conxuntos de datos de depresión de eRisk 2017, 2018 e 2020, analizando un total de 1.707 usuarios. Os resultados demostran que BERTopic supera significativamente a LDA e TopClus na avaliación humana da coherencia

temática (puntuacións medias: 2,49 fronte a 1,14 fronte a 1,61 nunha escala de 3 puntos, cun acordo inter-anotador $\kappa \approx 0,50$). A vantaxe de BERTopic provén do seu uso de transformadores de frases pre-adestrados, que manexan a variabilidade léxica do texto informal de forma máis efectiva que as representacións de bolsa de palabras de LDA. Cómpre sinalar que LDA foi avaliado con hiperparámetros por defecto, o que podería subestimar o seu potencial baixo un axuste coidadoso.

O estudo tamén demostra que as métricas automáticas de coherencia estándar correlacionan escasamente cos xuízos humanos sobre a calidade dos temas neste dominio. Nos datos de saúde mental de eRisk, NPMI acadou unha correlación de Spearman de só 0,112 cos xuízos humanos, mentres que UMass (0,001), UCI (0,04) e C_V (0,05) foron esencialmente non correlacionadas. Este fracaso é específico do dominio: as mesmas métricas acadan correlacións razoables en coleccións estándar, pero a linguaxe informal e emocionalmente cargada das redes sociais de saúde mental diverxe substancialmente dos textos para os que estas métricas foron deseñadas. Estes resultados motivaron o desenvolvemento de protocolos de avaliación humana e ferramentas de anotación. Máis alá da comparación de modelos, os temas descubertos demostraron relevancia clínica: a análise mediante BDI-II revelou fortes asociacións entre temas específicos (autolesións, dificultades nas relacións, perda de peso) e indicadores de depresión, e unha análise temporal capturou o incremento no discurso relacionado coa depresión durante o período pandémico de 2020-2021.

A terceira publicación aborda directamente o desafío da avaliación mediante un estudo exhaustivo de modelos de word embeddings para medir a coherencia temática. As métricas clásicas baseadas en co-ocorrencia (NPMI, UCI) compáranse con alternativas baseadas en embeddings empregando modelos estáticos (Word2Vec, GloVe, FastText) e modelos contextuais (BERT, RoBERTa, ALBERT, MPNET). O estudo avalíase sobre tres conxuntos de datos estándar con anotacións humanas de coherencia (20 Newsgroups, New York Times, Genomics), cada un con 100 temas de LDA puntuados por 24-26 anotadores nunha escala de tres puntos. O estudo tamén avalía o impacto da selección do corpus de referencia, demostrando que o rendemento de NPMI mellora drasticamente ao pasar do corpus orixinal a Wikipedia (de 0,581 a 0,735 en correlación de Spearman), confirmando que as métricas clásicas non miden unha propiedade intrínseca dos temas senón unha propiedade conxunta dos temas e do corpus de referencia.

Os resultados revelan que os embeddings de GloVe (con produto escalar) acadan unha correlación media de Spearman de 0,797 cos xuízos humanos de coherencia, superando a

mellor métrica clásica (NPMI con Wikipedia: 0,735). O rendemento de GloVe foi consistente nos tres conxuntos de datos, con correlacións superiores a 0,74 en todos os casos, un nivel de robustez inter-dominio que ningunha métrica clásica igualou. TBuckets, un método baseado en descomposición SVD e programación lineal enteira, acadou unha correlación lixeiramente superior (0,806), ofrecendo unha perspectiva alxébrica complementaria.

Un dos achados máis sorprendentes foi o escaso rendemento dos embeddings contextuais para esta tarefa. BERT acadou unha correlación media negativa ($-0,131$) porque a súa arquitectura require contexto a nivel de frase para producir representacións significativas; cando as palabras se presentan de forma illada –como na avaliación de coherencia– o modelo carece da entrada para a que foi deseñado. Un estudo de ablación a través das capas do transformador confirmou esta interpretación: o rendemento de BERT melloraba nas capas intermedias (2-5), onde o modelo comezou a integrar información posicional e estrutural pero aínda non se especializou na desambiguación contextual. Este achado establece unha recomendación práctica clara: para tarefas semánticas a nivel de palabra, os embeddings estáticos seguen sendo preferibles aos modelos contextuais. Esta recomendación ten un valor operativo directo para os profesionais que constrúen pipelines de modelado de temas, xa que os vectores de GloVe están pre-calculados, son de libre acceso e poden aplicarse sen reentrenar nin dispor de corpus específicos de dominio.

A cuarta publicación pasa da análise temática ao modelado temporal, propoñendo un marco novidoso baseado en word embeddings temporais para a detección temperá de signos de trastornos psicolóxicos. Mediante o seguimento de como o contexto semántico de palabras clave relacionadas coa saúde mental evoluciona ao longo do tempo nos historiais de publicación dos usuarios, o enfoque captura cambios lingüísticos indicativos de condicións emerxentes. O método emprega dous enfoques de embeddings temporais: TWEC, baseado en Word2Vec, que readestra un modelo para cada período temporal; e DCWE, baseado en BERT, que xera embeddings contextuais que son promediados por período. Para cada usuario, calcúlanse características delta que miden a magnitude do desprazamento de cada palabra entre o seu embedding temporal e un embedding de referencia. Estas características aliméntanse a clasificadores (SVM, Random Forest) para a detección.

Avaliado sobre 18 conxuntos de datos de eRisk que abarcan catro trastornos (depresión, anorexia, autolesións e xogo), o método demostra un rendemento competitivo na detección temperá, destacando especialmente na detección do xogo (F1 = 0,84 con DCWE-SVM en eRisk 2023) e producindo resultados notables en autolesións e anorexia. A análise de palabras

clave revelou que os usuarios positivos presentan movementos de palabras máis pronunciados e irregulares ca os usuarios de control, e que as sinaturas temporais son específicas de cada trastorno. Para o xogo, isto manifestouse como flutuacións dramáticas en termos directamente relacionados co trastorno, cunha correlación do termo máis relevante de 0,584. Para a depresión, os patróns foron máis difusos, distribuídos a través de vocabulario emocional xeral.

Non obstante, o enfoque presenta limitacións importantes. Os métodos de embeddings temporais foron deseñados orixinalmente para o cambio semántico diacrónico e son reutilizados aquí para unha tarefa onde o sinal relevante é o cambio emocional-contextual en usuarios individuais. Un usuario deprimido non emprega a palabra "durmir" nun novo sentido; emprégaa nun contexto angustioso. O marco captura este cambio só de forma indirecta, o que explica por qué funciona ben para o xogo (onde se introduce vocabulario específico do dominio) pero peor para a depresión (F1 = 0,35), onde o cambio reside no ton emocional das palabras comúns. Este contraste pon de relevo unha asimetría importante na detectabilidade lingüística das condicións de saúde mental e suxire que ningunha escolla arquitectónica única será suficiente ao longo de todo o espectro de trastornos. Ademais, o pipeline presenta un paradoxo denso-a-disperso: embeddings densos e custosos redúcense a valores escalares delta, descartando a información direccional do movemento semántico.

A quinta publicación presenta LabChain, un framework en Python para a experimentación científica reproducible e modular. Neste framework adóptase unha arquitectura de filtros e tubaxes onde cada paso de procesamento encapsula unha única transformación, e os filtros compóñense en pipelines serializables. Os atributos públicos (parámetros de configuración) restríxense a tipos serializables básicos, asegurando que calquera pipeline pode describirse completamente como un documento JSON. LabChain implementa un almacenamento en caché baseado en hash determinístico: a clave de caché de cada filtro calcúlase a partir do seu nome de clase, a súa configuración pública e, para filtros adestrados, o hash dos seus datos de adestramento, creando unha cadea de procedencia que identifica de forma única cada resultado intermedio.

A súa aplicación para reimplementar o pipeline de embeddings temporais da cuarta publicación revelou un erro de preprocesamento (un fallo ao aplanar datos xerárquicos) que permanecera oculto no código monolítico orixinal. A corrección deste erro, combinada coa exploración sistemática posibilitada polo almacenamento en caché, produciu melloras de rendemento de ata o 192,3 % nalgunhas tarefas. O almacenamento en caché eliminou aproximadamente 48 horas de computación redundante nos experimentos de detección de saúde mental,

correspondentes a aproximadamente 2,50 kg de emisións de CO_2. LabChain publícase como software de código aberto e emprégase en fluxos de traballo de investigación en produción no CiTIUS.

Respecto ás conclusións principais, a comparación de agrupamentos demostrou que os algoritmos non supervisados poden identificar estrutura temática en publicacións de Reddit relacionadas coa saúde mental, pero non separaron de forma efectiva as categorías diagnósticas, confirmando que a depresión non se manifesta como unha clase xeometricamente separable. O achado máis relevante foi que as métricas de validación interna non son comparables entre espazos de representación con propiedades xeométricas diferentes, un modo de fallo que probablemente afecte a calquera estudo comparativo de agrupamentos entre espazos de embeddings heteroxéneos. No ámbito do modelado de temas, BERTopic superou significativamente a LDA e TopClus na avaliación humana sobre datos de redes sociais, principalmente pola súa robustez á variabilidade léxica mediante transformadores de frases pre-adestrados. En canto á avaliación da coherencia, a métrica baseada en GloVe acadou un rendemento superior e consistente entre dominios en comparación con todas as métricas clásicas, mentres que os embeddings contextuais renderon escasamente porque requiren contexto a nivel de frase que está ausente na avaliación de coherencia a nivel de palabra. O enfoque de embeddings temporais demostrou que as dinámicas lingüísticas temporais portan un valor diagnóstico xenuíno, acadando resultados competitivos especialmente para a detección do xogo, pero o marco de cambio semántico diacrónico captura os cambios emocionais-contextuais só de forma indirecta, o que limita a súa efectividade para condicións como a depresión, onde o cambio reside no ton emocional e non na introdución de vocabulario. Finalmente, a reimplementación modular con LabChain expuxo un erro de preprocesamento oculto e permitiu unha exploración sistemática que produciu melloras de rendemento de ata o 192,3 %, demostrando que as prácticas de enxeñaría de software amplifican directamente a calidade científica.

A través das cinco publicacións emerxeu un achado transversal recorrente: a avaliación non é separable da representación. As métricas internas de agrupamento dependen da xeometría do espazo de embeddings. As métricas de coherencia dependen do corpus de referencia. As características temporais dependen do vocabulario de adestramento. En todos os casos, a fiabilidade da métrica está condicionada por propiedades do contexto representacional que a miúdo son invisibles para o investigador. Recoñecer e ter en conta estas dependencias é esencial para extraer conclusións válidas da análise computacional de texto. En síntese, cada métrica leva consigo unha teoría implícita do que se está a medir, e esa teoría raramente é neutral

respecto ás decisións tomadas augas arriba no pipeline.

Estes achados teñen implicacións para diversas áreas. Para a recuperación de información, a dependencia xeométrica das métricas internas de agrupamento implica que os estudos comparativos entre diferentes espazos de representación requiren validación externa. Para a análise de saúde mental, a detectabilidade dunha condición depende en parte da súa distintividade lingüística: as condicións con vocabulario específico e focalizado son máis susceptibles de detección a nivel de palabra ca as condicións con sinaturas lingüísticas difusas, o que suxire que a detección da depresión podería requirir métodos que operen a nivel de frase ou documento. Para a avaliación de modelos de temas, a avaliación humana segue sendo necesaria en dominios especializados, e a coherencia baseada en GloVe ofrece unha alternativa práctica, pre-computada e robusta entre dominios ás métricas clásicas. Para a investigación reproducible, o impacto de LabChain nesta tese ilustra que a modularización, o almacenamento en caché e a serialización non son meras conveniencias senón que poden mellorar directamente os resultados científicos. Máis alá destes achados específicos, o traballo defende un cambio que abandone as comparacións puntuais de modelos en favor de estudos sistemáticos que teñan en conta de forma explícita a interacción entre representación, avaliación e propiedades do conxunto de datos.

Varias direccións concretas de investigación emerxen deste traballo. A lagoa máis inmediata é a ausencia de bancos de proba de coherencia temática anotados por humanos sobre datos de redes sociais de saúde mental, o que permitiría a avaliación directa das métricas de coherencia baseadas en embeddings no dominio que motivou o seu desenvolvemento. As limitacións do enfoque diacrónico para a detección de saúde mental apuntan cara a arquitecturas que rastrexen traxectorias emocionais directamente sobre representacións densas, pasando de preguntar "como cambiou o significado desta palabra?" a "como evolucionou o ton emocional deste usuario?". Unha comparación sistemática con LDA axustado establecería de forma máis definitiva o alcance da vantaxe de BERTopic. A participación de psicólogos clínicos na avaliación de temas e traxectorias temporais reforzaría a ponte entre a análise computacional e a utilidade clínica. Finalmente, a extensión dos métodos a outras linguas, plataformas e contextos culturais é necesaria para avaliar a súa xeneralizabilidade. Igualmente importante, implicar aos clínicos non só como avaliadores senón tamén como co-deseñadores do pipeline de modelado podería axudar a aliñar as prioridades computacionais coas preguntas que máis importan na práctica clínica e asegurar que os avances nas métricas de detección se traduzan en melloras significativas na atención.

Summary

Mental health disorders constitute one of the greatest public health challenges of the twenty-first century. According to the World Health Organization, over 280 million people worldwide suffer from depression, and the prevalence of conditions such as anxiety, eating disorders, self-harm, and pathological gambling continues to rise. Early detection and timely intervention are widely recognized as critical factors in improving clinical outcomes; however, traditional diagnostic pathways remain limited by resource constraints, social stigma, and the inherent difficulty of identifying psychological distress before it escalates into a clinical crisis. The economic burden, estimated to exceed one trillion dollars annually in lost productivity worldwide, compounds the human cost and underscores the urgency of developing scalable, complementary tools that can support traditional clinical pathways without replacing them.

At the same time, social media platforms have become spaces where users disclose personal experiences with remarkable candor, often shielded by the perceived anonymity these platforms afford. Users frequently discuss their struggles with depression, eating disorders, and other mental health conditions in ways that would be difficult to capture through conventional clinical instruments. Research in psycholinguistics has established that language patterns correlate with psychological states, and social media provides an unprecedented window into these patterns at scale. The convergence of these two phenomena –the mental health crisis and the proliferation of digital self-expression– has given rise to a rapidly growing research area at the intersection of Natural Language Processing (NLP), Information Retrieval (IR), and Computational Psychology. Initiatives such as the eRisk shared tasks at CLEF and the CLPsych workshops have catalyzed community efforts by providing standardized datasets, evaluation protocols, and competitive benchmarks for the early detection of mental health risks from social media text. This opportunity, however, carries substantial ethical and methodological responsibilities, particularly around privacy preservation, the avoidance of stigmatizing

framings, and the recognition that computational signals must inform—rather than supplant—professional clinical judgment. However, the field continues to face fundamental challenges in text representation, pattern discovery, model evaluation, and experimental reproducibility.

This thesis, presented as a compendium of five publications, develops and evaluates computational methods for the analysis and early detection of mental health conditions from social media text, contributing advances in text clustering, topic modeling, coherence evaluation, temporal word embeddings, and reproducible experimentation. The work is structured around five research hypotheses, each addressed by a specific publication. The primary data source throughout the thesis is the collection of datasets released by the eRisk initiative, comprising posts and comments collected from Reddit, where users have been identified as exhibiting signs of specific mental health conditions or as belonging to a control group. Four mental health conditions are covered: depression (eRisk editions 2017, 2018, and 2022), anorexia (2018 and 2019), self-harm (2020 and 2021), and gambling (2022 and 2023). These datasets present characteristics that make them particularly challenging for computational analysis: the class distribution is heavily imbalanced, posts are written in informal English with significant variation in length and style, and users' posting histories span extended periods that introduce variability and seasonal factors. On average, each user contributes hundreds of posts distributed across months or years, providing a longitudinal trace of linguistic behavior that would be difficult to obtain through any other means. The text representations employed throughout the thesis range from classical sparse methods (bag-of-words, TF-IDF) to dense static embeddings (Word2Vec, GloVe, FastText) and contextual embeddings from transformer models (BERT, RoBERTa, ALBERT, MPNET). Applying clustering and topic modeling algorithms to this type of data faces the curse of dimensionality, which has led to the widespread adoption of pipeline architectures that decompose the task into stages: text representation, dimensionality reduction (using methods such as UMAP), and clustering or modeling on the reduced representations.

The first publication conducts a systematic comparison of clustering algorithms –partition-based (K-Means), model-based (Gaussian Mixture Models, GMM), and density-based (DBSCAN)– for knowledge discovery in Reddit data related to depression, using the eRisk 2017 dataset. The study employs multiple text representations, including Term Frequency (TF) vectors, TF-IDF, and contextual embeddings from transformer models (BERT, RoBERTa, GPT-2). Standard preprocessing including tokenization, stop-word filtering, and length-based selection was applied prior to representation, and the hyperparameters of each clustering algorithm

were explored across reasonable ranges to ensure that the observed differences reflected the underlying methods rather than incidental tuning choices. The results revealed that the best configuration –GMM with RoBERTa embeddings– discovered clusters corresponding to personal relationships, video games, and substance-related discussions, with the relationship cluster showing a higher proportion of depressed users. However, clustering did not effectively separate diagnostic categories (ARI = 0.180, NMI = 0.200), confirming that depression does not manifest as a geometrically separable class in any of the embedding spaces tested.

The most significant finding of this publication was methodological: internal validation metrics (Silhouette Coefficient, Calinski-Harabasz, Davies-Bouldin) are not comparable across representation spaces with different geometric properties. In our experiments, TF vectors –whose magnitude grows with document length– consistently maximized internal metrics while performing worst on external validation. Naïve reliance on internal metrics would have led to the selection of the worst representation method, illustrating a failure mode that is likely to affect any study comparing clusterings across heterogeneous embedding spaces. The implications extend well beyond mental health analysis, affecting any application in which clustering quality is assessed across heterogeneous representations, including recommender systems, document organization, and biomedical text mining.

The limitations of the clustering approach motivated a natural evolution toward topic modeling. Both techniques share a common computational backbone –embedding-based representation, dimensionality reduction, and grouping– but topic modeling adds a description layer that extracts representative terms for each group, transforming opaque clusters into interpretable topics. The second publication explores this direction, comparing LDA, BERTopic, and TopClus on the eRisk 2017, 2018, and 2020 depression datasets, analyzing a total of 1,707 users. The results demonstrate that BERTopic significantly outperforms LDA and TopClus in human evaluation of topic coherence (average ratings: 2.49 vs. 1.14 vs. 1.61 on a 3-point scale, with inter-annotator agreement $\kappa \approx 0.50$). BERTopic's advantage stems from its use of pre-trained sentence transformers, which handle the lexical variability of informal text more effectively than LDA's bag-of-words representations. It should be noted that LDA was evaluated with default hyperparameters, which may underestimate its potential under careful tuning.

The study also demonstrates that standard automatic coherence metrics correlate poorly with human judgments of topic quality in this domain. On the eRisk mental health data, NPMI achieved a Spearman correlation of only 0.112 with human judgments, while UMass (0.001),

UCI (0.04), and C_V (0.05) were essentially uncorrelated. This failure is domain-specific: the same metrics achieve reasonable correlations on standard benchmarks, but the informal, emotionally charged vocabulary of mental health social media diverges substantially from the text these metrics were designed for. These results motivated the development of human evaluation protocols and annotation tools. Beyond model comparison, the discovered topics proved clinically relevant: BDI-II analysis revealed strong associations between specific topics (self-harm, relationship difficulties, weight loss) and depression indicators, and a temporal analysis captured the increase in depression-related discourse during the 2020–2021 pandemic period.

The third publication addresses the evaluation challenge directly by conducting a comprehensive study of word embedding models for measuring topic coherence. Classical co-occurrence metrics (NPMI, UCI) are compared against embedding-based alternatives using static models (Word2Vec, GloVe, FastText) and contextual models (BERT, RoBERTa, ALBERT, MPNET). The study is evaluated on three standard datasets with human coherence annotations (20 Newsgroups, New York Times, Genomics), each containing 100 LDA topics rated by 24–26 human annotators on a three-point scale. The study also evaluates the impact of reference corpus selection, demonstrating that NPMI's performance improves dramatically when switching from the original corpus to Wikipedia (from 0.581 to 0.735 in Spearman correlation), confirming that classical metrics do not measure an intrinsic property of topics but rather a joint property of the topics and the reference corpus.

The results reveal that GloVe embeddings (with dot product similarity) achieve an average Spearman correlation of 0.797 with human coherence judgments, surpassing the best classical metric (NPMI with Wikipedia: 0.735). GloVe's performance was consistent across all three datasets, with correlations exceeding 0.74 in every case –a level of cross-domain robustness that no classical metric matched. TBuckets, a method based on SVD decomposition and integer linear programming, achieved slightly higher correlation (0.806), offering a complementary algebraic perspective.

One of the most surprising findings was the poor performance of contextual embeddings for this task. BERT achieved a negative average correlation (-0.131) because its architecture requires sentence-level context to produce meaningful representations; when words are presented in isolation –as in coherence evaluation– the model lacks the input it was designed to exploit. An ablation study across transformer layers confirmed this interpretation: BERT's coherence performance improved at intermediate layers (layers 2–5), where the model has be-

gun integrating positional and structural information but has not yet specialized for contextual disambiguation. This finding establishes a clear practical guideline: for word-level semantic tasks, static embeddings remain preferable to contextual models. This recommendation has direct operational value for practitioners building topic modeling pipelines, since GloVe vectors are pre-computed, freely available, and can be applied without retraining or access to domain-specific corpora.

The fourth publication shifts from thematic analysis to temporal modeling, proposing a novel framework based on temporal word embeddings for the early detection of signs of psychological disorders. By tracking how the semantic context of mental health-related keywords evolves over time in users' posting histories, the approach captures linguistic shifts indicative of emerging conditions. The method employs two temporal embedding approaches: TWEC, based on Word2Vec, which retrains a model for each time period; and DCWE, based on BERT, which generates contextual embeddings that are averaged per period. For each user, delta features are computed that measure the magnitude of displacement of each word between its temporal embedding and a reference embedding. These features are fed to classifiers (SVM, Random Forest) for detection.

Evaluated on 18 eRisk datasets spanning four disorders –depression, anorexia, self-harm, and gambling– the method demonstrates competitive early detection performance, particularly excelling on gambling detection (F1 = 0.84 with DCWE-SVM on eRisk 2023) and producing respectable results on self-harm and anorexia. Keyword analysis revealed that positive users exhibit more pronounced and irregular word movements than controls, and that temporal signatures are disorder-specific. For gambling, this manifested as dramatic fluctuations in terms directly related to the disorder, with the top word correlation reaching 0.584. For depression, the patterns were more diffuse, spread across general emotional vocabulary.

However, the approach has important limitations. The temporal embedding methods were originally designed for diachronic semantic change and are repurposed here for a task where the relevant signal is emotional-contextual change in individual users. A depressed user does not use the word "sleep" in a new sense; they use it in a distressed context. The framework captures this shift only indirectly, which explains why it works well for gambling (where domain-specific vocabulary is introduced) but poorly for depression (F1 = 0.35), where the change lies in the emotional tone of common words. This contrast highlights an important asymmetry in the linguistic detectability of mental health conditions and suggests that no single architectural choice will suffice across the full spectrum of disorders. Additionally,

the pipeline exhibits a dense-to-sparse paradox: expensive dense embeddings are reduced to scalar delta values, discarding the directional information of semantic movement.

The fifth publication presents LabChain, a Python framework for reproducible and modular scientific experiments. The framework adopts a pipes-and-filters architecture where each processing step encapsulates a single transformation, and filters are composed into serializable pipelines. Public attributes (configuration parameters) are restricted to basic serializable types, ensuring that any pipeline can be fully described as a JSON document. LabChain implements deterministic hash-based caching: each filter's cache key is computed from its class name, public configuration, and –for trainable filters– the hash of its training data, creating a provenance chain that uniquely identifies every intermediate result.

Its application to re-implement the temporal embedding pipeline from the fourth publication revealed a preprocessing bug –a failure to flatten hierarchical data– that had remained hidden in the original monolithic code. Fixing this bug, combined with the systematic exploration enabled by caching, produced performance improvements of up to 192.3% on some tasks. Caching eliminated approximately 48 hours of redundant computation across the mental health detection experiments, corresponding to approximately 2.50 kg of CO_2 emissions. LabChain is published as open-source software and is used in production research workflows at CiTIUS.

Regarding the main conclusions, the clustering comparison demonstrated that unsupervised algorithms can identify thematic structure in mental health-related Reddit posts, but did not separate diagnostic categories effectively, confirming that depression does not manifest as a geometrically separable class. The most consequential finding was that internal validation metrics are not comparable across representation spaces with different geometric properties, a failure mode likely to affect any comparative clustering study across heterogeneous embedding spaces. In the area of topic modeling, BERTopic significantly outperformed LDA and TopClus in human evaluation on social media data, primarily due to its robustness to lexical variability through pre-trained sentence transformers. Regarding coherence evaluation, GloVe-based metrics achieved superior and consistent cross-domain performance compared to all classical metrics, while contextual embeddings performed poorly because they require sentence-level context that is absent in word-level coherence evaluation. The temporal embedding approach demonstrated that temporal linguistic dynamics carry genuine diagnostic value, achieving competitive results particularly for gambling detection, but the diachronic semantic change framework captures emotional-contextual shifts only indirectly, which limits its effectiveness

for conditions like depression where the change lies in emotional tone rather than vocabulary introduction. Finally, LabChain's modular re-implementation exposed a hidden preprocessing bug and enabled systematic exploration that produced performance improvements of up to 192.3%, demonstrating that software engineering practices directly amplify scientific quality.

Across all five publications, a recurring transversal finding emerged: evaluation is not separable from representation. Internal clustering metrics depend on the geometry of the embedding space. Coherence metrics depend on the reference corpus. Temporal features depend on the training vocabulary. In every case, the metric's reliability is contingent on properties of the representational context that are often invisible to the practitioner. Recognizing and accounting for these dependencies is essential for drawing valid conclusions from computational text analysis. In short, every metric carries an implicit theory of what is being measured, and that theory is rarely neutral with respect to the choices made upstream in the pipeline.

These findings carry implications for several areas. For information retrieval, the geometry-dependence of internal clustering metrics implies that comparative studies across different representation spaces require external validation. For mental health analysis, the detectability of a condition depends partly on its linguistic distinctiveness: conditions with focused, domain-specific vocabulary are more amenable to word-level detection than conditions with diffuse linguistic signatures, suggesting that depression detection may require methods operating at the sentence or document level. For topic model evaluation, human evaluation remains necessary in specialized domains, and GloVe-based coherence offers a practical, pre-computed, cross-domain robust alternative to classical metrics. For reproducible research, LabChain's impact on this thesis illustrates that modularization, caching, and serialization are not mere conveniences but can directly improve scientific outcomes. Beyond these specific findings, the work argues for a shift away from one-shot model comparisons toward systematic studies that explicitly account for the interaction between representation, evaluation, and dataset properties.

Several concrete research directions emerge from this work. The most immediate gap is the absence of human-annotated topic coherence benchmarks on mental health social media data, which would allow direct evaluation of embedding-based coherence metrics in the domain that motivated their development. The limitations of the diachronic approach for mental health detection point toward architectures that track emotional trajectories directly on dense representations, shifting from asking "how has this word's meaning changed?" to "how has this user's emotional tone evolved?". A systematic comparison with properly tuned LDA

would establish more definitively the extent of BERTopic's advantage. Involving clinical psychologists in the evaluation of topics and temporal trajectories would strengthen the bridge between computational analysis and clinical utility. Finally, extending the methods to other languages, platforms, and cultural contexts is necessary to assess their generalizability. Equally important, engaging clinicians not only as evaluators but also as co-designers of the modeling pipeline could help align computational priorities with the questions that matter most in clinical practice and ensure that progress in detection metrics translates into meaningful improvements in care.

Chapter 1

Introduction

Mental health disorders represent one of the most pressing public health challenges of the twenty-first century. According to the World Health Organization, over 280 million people worldwide suffer from depression [66], and the prevalence of conditions such as anxiety, eating disorders, self-harm, and pathological gambling continues to rise [32, 54]. The consequences of these disorders extend far beyond individual suffering, placing an enormous burden on healthcare systems, economies, and social structures. Early detection and timely intervention are widely recognized as critical factors in improving outcomes, yet traditional diagnostic pathways remain limited by resource constraints, social stigma, and the inherent difficulty of identifying psychological distress before it escalates into a clinical crisis [20].

Simultaneously, the digital transformation of social interaction has fundamentally altered the way individuals express themselves, seek support, and construct their identities. Social media platforms –including Reddit, Twitter, and various online forums– have become spaces where users disclose personal experiences with remarkable candor, often shielded by the perceived anonymity these platforms afford [23]. This openness extends to sensitive domains: users frequently discuss their struggles with depression, suicidal ideation, eating disorders, and other mental health conditions in ways that would be difficult to capture through conventional clinical instruments [64, 61]. Research in psycholinguistics has long established that language patterns correlate with psychological states [52], and social media provides an unprecedented window into these patterns at scale.

The convergence of these two phenomena –the mental health crisis and the proliferation of digital self-expression– has given rise to a rapidly growing research area at the intersec-

tion of Natural Language Processing (NLP), Information Retrieval (IR), and Computational Psychology (CP). Initiatives such as the eRisk shared tasks at CLEF [36, 37, 47, 48, 14] and the CLPsych workshops [15, 43, 38, 69] have catalyzed community efforts by providing standardized datasets, evaluation protocols, and competitive benchmarks for the early detection of mental health risks from social media text. These efforts have demonstrated that computational approaches can indeed identify individuals at risk, but the field continues to face fundamental challenges in text representation, pattern discovery, model evaluation, and experimental reproducibility.

The sheer volume of user-generated content on social media platforms constitutes an enormous, continuously growing corpus of naturalistic language data. Every day, millions of posts, comments, and interactions are produced, each carrying traces of the author's emotional state, cognitive patterns, and behavioral tendencies. From a computational perspective, this data represents an exploitable resource of extraordinary richness –one that, if properly analyzed, can yield insights into human behavior. This new form of screening can complement and extend traditional psychological assessment methods [5, 68].

The challenge, however, lies in transforming this raw, unstructured, and inherently noisy data into actionable knowledge. Social media text is characterized by informal language, abbreviations, slang, code-switching, irony, and highly variable post lengths [7]. Users may discuss multiple topics within a single post, employ sarcasm that inverts the literal meaning of their words, or express distress through subtle linguistic shifts rather than explicit statements. These characteristics make social media data particularly resistant to analysis by methods originally designed for well-edited, domain-specific corpora.

The goal of extracting meaningful patterns from social media data requires a diverse arsenal of computational techniques, each offering different perspectives on the underlying linguistic phenomena. At the highest level, these techniques can be organized along two fundamental axes: the distinction between supervised and unsupervised learning, and the distinction between static and temporal analysis.

1.1 Unsupervised Methods: Discovering Structure Without Labels

Unsupervised methods are particularly valuable in the mental health domain because labeled data is scarce, expensive to produce, and ethically sensitive to collect. These methods aim

to discover latent structure in data without relying on predefined categories, making them well-suited for exploratory analysis where the research question is not "does this user belong to category X?" but rather "what meaningful groups and patterns exist in this population?".

Clustering algorithms constitute one of the most fundamental families of unsupervised methods for text analysis. Given a collection of documents represented as vectors in some feature space, clustering seeks to partition the collection into groups such that documents within the same group are more similar to each other than to documents in other groups [30, 29]. A variety of algorithmic paradigms have been proposed for this task, each embodying different assumptions about the shape and structure of clusters:

- **Partition-based methods**, such as K-Means [39], iteratively assign data points to the nearest cluster centroid and update centroids until convergence. Their computational efficiency makes them a natural default for large-scale text collections, but they assume spherical, equally-sized clusters and require the number of clusters to be specified in advance.

- **Model-based methods**, such as Gaussian Mixture Models (GMM) [24, 59], treat each cluster as a probability distribution and assign data points to clusters based on maximum likelihood. They offer greater flexibility than partition-based methods by allowing elliptical cluster shapes and soft assignments, but their performance degrades in high-dimensional spaces.

- **Hierarchical methods**, including agglomerative approaches such as Ward's method [22], build a tree-like structure of nested clusters that can be cut at different levels to produce partitions of varying granularity. They provide interpretable dendrograms but are computationally expensive for large datasets.

- **Density-based methods**, such as DBSCAN [28] and OPTICS [4], define clusters as regions of high density separated by regions of low density. They can discover clusters of arbitrary shape and naturally identify outliers, but their performance is sensitive to the density parameters and they struggle with clusters of varying density.

Beyond clustering, other unsupervised paradigms have proven valuable for text analysis. Masked language modeling, as exemplified by BERT [25] and its variants, learns rich contextual representations by predicting masked tokens in their surrounding context. These

self-supervised methods produce dense vector representations that encode semantic and syntactic information, and have become the foundation for a wide range of downstream NLP applications.

1.1.1 The Curse of Dimensionality in Text Clustering

A fundamental obstacle in applying clustering algorithms to text data is the curse of dimensionality. Traditional text representations, such as bag-of-words or TF-IDF vectors, produce extremely high-dimensional and sparse feature spaces where the number of dimensions (vocabulary terms) can easily reach tens or hundreds of thousands. In such spaces, the distance metrics upon which most clustering algorithms rely become increasingly uninformative: as dimensionality grows, the relative difference between the nearest and farthest points from any given query tends to zero, effectively rendering distance-based methods unable to distinguish meaningful structure from noise [30].

This fundamental limitation has led to the widespread adoption of pipeline architectures that decompose the clustering task into two stages: first, a dimensionality reduction step that projects the high-dimensional text representations into a lower-dimensional space that preserves the most salient structure; and second, a clustering step that operates on the reduced representations. The dimensionality reduction component may employ linear methods such as Principal Component Analysis (PCA) or Singular Value Decomposition (SVD), or nonlinear methods such as UMAP or t-SNE, depending on the nature of the data and the downstream clustering algorithm.

The choice of text representation itself introduces an additional layer of complexity. Traditional sparse representations (bag-of-words, TF-IDF) capture lexical co-occurrence but miss semantic relationships between words. Dense representations produced by word embedding models –including static embeddings such as Word2Vec [42], GloVe [53], and FastText [10], as well as contextual embeddings from transformer-based models such as BERT [25], RoBERTa, ALBERT, and MPNET– encode richer semantic information but operate in spaces with different geometric properties. This diversity of representation methods creates a landscape in which the interaction between representation, dimensionality reduction, and clustering algorithm becomes a critical design decision with significant impact on the quality of the discovered patterns.

1.1.2 Evaluating Cluster Quality: Metrics and Their Limitations

The evaluation of clustering results presents its own set of challenges, intimately connected to the unsupervised nature of the task. Two broad families of evaluation metrics have been developed, each with distinct assumptions and limitations.

Internal validation metrics assess clustering quality based solely on the intrinsic properties of the partition, without reference to any external ground truth. Metrics such as the Silhouette Coefficient [62], the Calinski-Harabasz Index [12], and the Davies-Bouldin Index [21] quantify the degree to which clusters are compact and well-separated. These metrics are appealing because they do not require labeled data, but they carry a critical limitation: they are defined in terms of distances computed in the feature space in which the clustering was performed. This means that internal metrics cannot be meaningfully compared across clusterings performed in different vector spaces. For instance, comparing a clustering obtained from TF-IDF representations with one obtained from BERT embeddings using the Silhouette Coefficient would be comparing values computed under fundamentally incompatible distance functions.

External validation metrics, such as the Rand Index [56], Normalized Mutual Information (NMI) [65], and the V-measure, evaluate clustering quality by measuring agreement between the discovered partition and a known ground truth labeling. These metrics allow meaningful comparison across different representations and algorithms, but they require a reliable ground truth, a condition that is frequently unmet in exploratory scenarios where the very purpose of clustering is to discover previously unknown structure.

This tension between the need for cross-method comparison and the limitations of available metrics constitutes a recurring theme throughout this thesis. It motivates not only the methodological choices made in the individual publications but also the development of alternative evaluation approaches, including embedding-based coherence metrics and human evaluation protocols.

1.1.3 Topic Models: From Clustering to Thematic Understanding

Topic models can be understood as a specialized evolution of the clustering paradigm, designed specifically for the discovery of thematic structure in text collections [8, 11, 2]. Rather than simply grouping documents into clusters, topic models aim to identify latent *topics* (distributions over words that explain the observed patterns in the corpus) and to characterize

each document as a mixture of these topics. This richer representational framework makes topic models particularly attractive for understanding the thematic concerns expressed in social media posts related to mental health [50, 46, 45].

The foundational model in this family is Latent Dirichlet Allocation (LDA) [9], a probabilistic generative model that assumes each document is generated by first sampling a distribution over topics, and then sampling words from the corresponding topic-word distributions. LDA and its extensions have been extensively applied across a wide range of domains and have become a standard tool in computational social science. However, LDA's assumptions –including the bag-of-words representation and the Dirichlet prior– impose limitations on its ability to capture the semantic nuances of modern text data.

More recent approaches have sought to address these limitations by integrating modern representation learning into the topic modeling pipeline. BERTopic [31] exemplifies this trend: it constructs topics by first encoding documents using pre-trained sentence transformers, then reducing dimensionality with UMAP, clustering the reduced representations with HDBSCAN, and finally extracting topic descriptions using a class-based TF-IDF procedure. This modular architecture –essentially a pipeline of representation, reduction, clustering, and description– makes explicit the connection between topic modeling and the clustering pipelines discussed earlier, while leveraging the superior representational capacity of modern language models. Similarly, TopClus [41] employs a latent category modeling approach that combines pre-trained neural language models with a joint objective for topic-level and document-level representations.

Despite their sophistication, topic models introduce their own evaluation challenges. The standard approach of assessing topic quality through automatic coherence metrics, such as NPMI and UCI, has been shown to correlate poorly with human judgments of topic interpretability [35, 27]. This finding has profound implications: if the metrics used to select and compare topic models do not reliably capture the property they are meant to measure, then the conclusions drawn from comparative studies based solely on these metrics may be misleading. Alternative evaluation strategies, including human evaluation and embedding-based coherence metrics, offer promising directions but bring their own costs and limitations.

1.2 Supervised Methods: Learning from Labeled Examples

When labeled data is available, supervised methods offer a complementary perspective. Fine-tuning pre-trained language models on task-specific datasets has become the dominant paradigm for text classification [1, 40], enabling models to leverage the linguistic knowledge acquired during pre-training while adapting to the specific vocabulary and patterns of the target domain. In the context of mental health detection, supervised classifiers can be trained to distinguish between users exhibiting signs of a particular disorder and control users, achieving increasingly competitive performance as models and datasets have matured.

Information retrieval techniques also play a role in this ecosystem, particularly in the framing of early risk detection as a sequential decision problem. Rather than classifying a complete user history at once, early detection systems must process posts incrementally as they arrive, deciding at each step whether sufficient evidence has accumulated to issue an alert [36]. This temporal, sequential dimension introduces unique challenges that static classification models are ill-equipped to address.

1.3 The Temporal Dimension: Language Change as a Signal

A distinguishing characteristic of mental health analysis in social media is the importance of the temporal dimension. Unlike static classification tasks where a single snapshot of a user's language is used to make a prediction, early risk detection requires monitoring how a user's language *evolves over time* [36]. A user who transitions from discussing everyday activities in positive terms to increasingly expressing hopelessness, social withdrawal, or preoccupation with death may be exhibiting early signs of a developing depressive episode. Capturing such transitions requires methods that are sensitive to temporal dynamics.

Temporal word embeddings provide a principled framework for modeling such linguistic change [33, 67, 63]. By training separate embedding spaces for different time periods and then aligning them to enable comparison, these methods can quantify how the meaning or usage context of specific words changes over time. When applied to the posts of individual social media users, temporal embeddings can detect shifts in the semantic context of mental health-related keywords –for instance, tracking whether the word "sleep" moves from a neutral context ("I slept well") to a distress-related context ("I can't sleep anymore") [26, 34]. Methods such as Temporal Word Embeddings with a Compass (TWEC) [26] and Dynamic Contextualized

Word Embeddings (DCWE) [34] have been proposed for this purpose, each offering different trade-offs between computational cost, alignment quality, and sensitivity to semantic change.

1.4 Reproducibility and Software Engineering in Research

The computational complexity of the methods discussed above –involving large pre-trained models, expensive embedding computations, multiple hyperparameter configurations, and iterative experimental cycles– places significant demands on research infrastructure. Reproducibility, widely regarded as a cornerstone of scientific practice, is particularly challenging in machine learning research, where results depend on specific software versions, random seeds, hardware configurations, and often on expensive computations that are difficult to replicate exactly [6].

A common practical problem in this type of research projects is the redundant recomputation of expensive intermediate results. When a research team explores different clustering algorithms on the same set of document embeddings, for instance, the embedding computation must be performed once but is frequently repeated due to inadequate caching or poor experiment management. This waste of resources is magnified in collaborative settings where multiple researchers work on related problems but lack mechanisms to share intermediate computations.

Existing frameworks address aspects of this problem: scikit-learn [51] provides standardized interfaces for machine learning components; workflow engines such as Snakemake [44] and Luigi [60] manage task dependencies in data processing pipelines; and tools such as PyCaret [3] and Kedro [55] offer higher-level abstractions for experiment management. However, none of these solutions fully address the need for automatic, content-addressable caching of intermediate results that enables transparent reuse across experiments and across teams. Such capability is essential for efficient and reproducible research at the scale demanded by modern NLP experimentation.

1.5 Publications

The research developed in this thesis has led to the following contributions:

Journals:

Couto, M.[a], Parapar, J.[b], and Losada, D. E.[a] (2024). *Comparison of Clustering Algorithms for Knowledge Discovery in Social Media Publications: A Case Study of Mental Health Analysis*. Procesamiento del Lenguaje Natural (SEPLN), 73. The publication is available at: `http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6601`

[a]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Spain.
[b]Information Retrieval Lab, CITIC, Universidade da Coruña, Spain.

- **Quality indicators**: SJR: 0.570 (2024), Q2 in Computer Science – Applications, Q1 in Linguistics and Language. Indexed in Scopus and DBLP.
- **PhD candidate contribution**: Research conceptualization, design and implementation of the experimental pipeline, clustering evaluation, analysis, and manuscript writing.
- **Reproduction rights**: Published under CC BY-NC 4.0 license. Reproduction in this thesis is permitted under the terms of the license.

Couto, M.[a], Parapar, J.[b], and Losada, D. E.[a] (2025). *Exploiting Topic Analysis Models to Explore Psychological Dimensions in Social Media Data*. Scientific Reports (Nature Portfolio). The publication is available at: `https://doi.org/10.1038/s41598-026-36339-y`

[a]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Spain.
[b]Information Retrieval Lab, CITIC, Universidade da Coruña, Spain.

- **Quality indicators**: JCR Impact Factor: 3.8 (2024). SJR: 0.874 (2024), Q1 in Multidisciplinary Sciences.
- **PhD candidate contribution**: Research conceptualization, design and implementation of the experimental framework, topic model evaluation, human evaluation protocol design, data curation, and manuscript writing.
- **Reproduction rights**: Published under CC BY 4.0 license (Springer Nature Open Access). Reproduction in this thesis is permitted under the terms of the license.

Couto, M.[a*], Pérez, A.[b*], Parapar, J.[b], and Losada, D. E.[a] (2025). *Temporal Word Embeddings for Early Detection of Psychological Disorders on Social Media*. Journal of Healthcare Informatics Research (JHIR), Springer, pp. 1–30. Published online: January 22, 2025.
[*]Equal contribution. The publication is available at: `https://doi.org/10.1007/s41666-025-00186-9`

[a]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Spain.
[b]Information Retrieval Lab, CITIC, Universidade da Coruña, Spain.

- **Quality indicators**: JCR Impact Factor: 5.4 (2024). SJR: 1.385 (2024), Q1 in Health Informatics.
- **PhD candidate contribution**: Research conceptualization, design and implementation of the temporal embedding framework, experimentation, keyword analysis, and manuscript writing (equal contribution with A. Perez).
- **Reproduction rights**: Published by Springer Nature. Reproduction in this thesis is permitted under the Springer Nature author rights policy for inclusion in doctoral dissertations.

Couto, M.[a], Parapar, J.[b], and Losada, D. E.[a] (2025). *LabChain: Enabling Reproducible and Modular Scientific Experiments in Python*. SoftwareX, Elsevier. The publication is available at: `https://doi.org/10.1016/j.softx.2026.102543`

[a]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Spain.
[b]Information Retrieval Lab, CITIC, Universidade da Coruña, Spain.

- **Quality indicators**: JCR Impact Factor: 2.7 (2024). SJR: 0.483 (2024), Q2 in Software.
- **PhD candidate contribution**: Research conceptualization, design and full implementation of the LabChain framework, experimentation, documentation, and manuscript writing.
- **Reproduction rights**: Published under CC BY-NC 4.0 license. Reproduction in this thesis is permitted under the terms of the license.

Journal publications under review:

Couto, M.[a], Parapar, J.[b], and Losada, D. E.[a] (2025). *A Study of Word Embedding Models for Measuring Topic Coherence*. Submitted to Knowledge and Information Systems (KAIS), Springer.

[a]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Spain.
[b]Information Retrieval Lab, CITIC, Universidade da Coruña, Spain.

- **Quality indicators**: JCR Impact Factor: 2.5 (2024). SJR: 0.827 (2024), Q2 in Computer Science – Information Systems.
- **PhD candidate contribution**: Research conceptualization, design and implementation of the evaluation framework, experimentation, analysis, and manuscript writing.
- **Reproduction rights**: Published by Springer Nature. Reproduction in this thesis is permitted under the Springer Nature author rights policy for inclusion in doctoral dissertations.

Other conference publications related to this thesis:

Couto, M.[a], Fernández-Pichel, M.[a], Aragon, M.E.[a], and Losada, D. E.[a] (2025). *Analyzing Gambling Addictions: A Spanish Corpus for Understanding Pathological Behavior*. In Findings of the Association for Computational Linguistics: EMNLP 2025, pp. 17610–17619, Suzhou, China. ACL. The publication is available at: `https://doi.org/10.18653/v1/2025.findings-emnlp.955`

[a]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Spain.

- **Quality indicators**: Conference ranked as Class 1 in GII-GRIN-SCIE (GGS) Conference Rating and also ranked as CORE:A*.
- **PhD candidate contribution**: Research conceptualization, corpus creation, experimentation, analysis, and manuscript writing.
- **Reproduction rights**: Published under CC BY 4.0 license (ACL Anthology). Reproduction in this thesis is permitted under the terms of the license.

1.6 Justification of Unity and Coherence of the Thesis

The five publications that constitute this thesis address the challenges outlined above from complementary perspectives, forming a coherent research program that progresses from exploratory analysis to sophisticated detection systems, supported by methodological contributions to evaluation and reproducible experimentation.

Publication I (Section 4.1) establishes the application domain by conducting a systematic comparison of clustering algorithms for knowledge discovery in mental health-related social media data. This work directly confronts the challenges of text representation, dimensionality reduction, and cluster evaluation discussed in Sections 1.1.1 and 1.1.2, providing empirical evidence on the relative strengths of partition-based, hierarchical, density-based, and model-based approaches when applied to the noisy, high-dimensional text characteristic of online mental health communities.

Publication II (Section 4.2) deepens and extends this exploratory analysis by moving from generic clustering to topic modeling. Motivated by the limitations of clustering-based approaches for thematic understanding, this work conducts an extensive comparison of topic analysis models –including LDA, BERTopic, and TopClus– applied to psychological dimensions in social media data. It demonstrates the superiority of BERTopic for noisy social media text, reveals the inadequacy of standard automatic coherence metrics, and introduces human evaluation protocols that provide more reliable assessments of topic quality.

Publication III (Section 5.1) addresses a fundamental methodological question raised by the first two publications: how should topic quality be evaluated? By conducting a comprehensive study of word embedding models for measuring topic coherence, this work provides a systematic evaluation of embedding-based alternatives to classical co-occurrence metrics. It demonstrates that embedding-based coherence measures, particularly those derived from certain layers and configurations of pre-trained models, can offer competitive or superior alignment with human judgments, thereby providing the evaluation foundations upon which the topic modeling work in Publication II relies.

Publication IV (Section 4.3) shifts from thematic analysis to temporal modeling, introducing temporal word embeddings for the early detection of psychological disorders from social media. Building upon the embedding expertise developed throughout the preceding publications, this work demonstrates that tracking linguistic change over time through temporal embedding alignment provides competitive performance for early risk detection across multiple mental health conditions. It represents the culmination of the progression from static,

exploratory analysis to dynamic, detection-oriented systems.

Publication V (Chapter 4.4) provides the software engineering foundation that enabled and improved the experimental work across the research program. The LabChain framework, designed for reproducible and modular scientific experiments, addresses the computational efficiency and reproducibility challenges discussed in Section 1.4. Its hash-based caching mechanism eliminates redundant computation of expensive embeddings and transformations, while its pipeline architecture enforces the modular experimental designs employed throughout the preceding publications.

Together, these five publications trace a coherent research program from understanding the data (clustering and topic modeling), through developing better evaluation tools (embedding-based coherence metrics), to building more sophisticated analysis systems (temporal embeddings for early detection), all supported by a purpose-oriented software infrastructure (LabChain). This thesis thus contributes not only individual results in each of these areas but also a unified perspective on how these components interact in the pursuit of computational approaches to mental health analysis from social media.

CHAPTER 2

HYPOTHESES AND OBJECTIVES

This chapter formulates the general research hypotheses and objectives that guided the work presented in this thesis. The hypotheses are derived from the open questions and challenges identified in the introduction, and each is addressed by one or more of the publications that constitute the compendium. The objectives translate these hypotheses into specific research goals.

2.1 Hypotheses

The following hypotheses articulate the core assumptions and research propositions that this thesis investigates:

H1. *Clustering algorithms can uncover meaningful structure in mental health-related social media data.* Social media posts authored by individuals with mental health conditions contain latent thematic patterns that can be discovered through unsupervised clustering. Different families of clustering algorithms –partition-based, hierarchical, density-based, and model-based– exhibit different strengths and weaknesses when applied to this type of noisy, high-dimensional textual data, and their systematic comparison can reveal which approaches are most suitable for psychological profiling.

H2. *Neural topic models produce more coherent and interpretable topics than classical generative models on noisy social media text.* Modern topic modeling approaches that incorporate pre-trained language model representations, such as BERTopic, generate topics of higher quality than classical probabilistic models such as LDA when applied to

the informal, sparse, and heterogeneous text characteristic of social media. Furthermore, automatic coherence metrics provide an incomplete picture of topic quality, and human evaluation is necessary to reliably assess the interpretability and clinical relevance of the discovered topics.

H3. *Word embedding-based metrics offer a competitive alternative to classical co-occurrence metrics for evaluating topic coherence.* Coherence metrics based on word embedding similarity –exploiting semantic representations learned by models such as Word2Vec, GloVe, FastText, and transformer-based architectures– can match or surpass the correlation with human judgments achieved by classical co-occurrence metrics such as NPMI and UCI. Moreover, the choice of embedding model and, in the case of transformer architectures, the specific layer from which representations are extracted, significantly affects the quality of the resulting coherence estimates.

H4. *Temporal word embeddings can capture linguistic changes indicative of emerging psychological disorders.* The progression of mental health conditions is accompanied by detectable changes in language use over time. By training temporal word embeddings on successive periods of a user's posting history and measuring the displacement of mental health-related keywords in the embedding space, it is possible to construct features that enable competitive early risk detection across multiple psychological disorders.

H5. *A modular pipeline framework with automatic caching can significantly improve the reproducibility and efficiency of text mining experiments.* A software framework based on pipeline architectures, hash-based content-addressable caching, and serializable experiment configurations can eliminate redundant computations across experiments and research teams, improve code quality through enforced modularity, and lead to measurable improvements in both computational efficiency and scientific outcomes.

2.2 General Objective

The overarching objective of this thesis is:

> *To develop and evaluate computational methods for the analysis and early detection of mental health conditions from social media text, advancing the state of the art in knowledge acquisition, text representation, thematic analysis, evaluation methodology, temporal modeling, and reproducible experimentation.*

This general objective encompasses two complementary dimensions: an *application dimension*, concerned with extracting actionable knowledge about mental health from user-generated content; and a *methodological dimension*, concerned with developing and rigorously evaluating the tools, metrics, and infrastructure required to conduct this type of research effectively and reproducibly.

2.3 Specific Objectives

The general objective is decomposed into the following specific objectives:

O1. *Compare and evaluate clustering algorithms for knowledge discovery in mental health social media data.* Conduct a systematic comparison of partition-based, hierarchical, density-based, and model-based clustering algorithms applied to Reddit posts from individuals with and without depression. Assess the suitability of each algorithm family for identifying thematic groups and community-level patterns correlated with mental health conditions, using both internal and external validation metrics.

O2. *Compare topic analysis models for exploring psychological dimensions in social media data.* Evaluate the performance of LDA, BERTopic, and TopClus for discovering psychologically relevant topics in eRisk depression datasets. Assess topic quality using both automatic coherence metrics and human evaluation, develop annotation tools and protocols for human assessment, and analyze the clinical relevance of the discovered topics (including temporal effects, such as the influence of the COVID-19 pandemic on mental health discourse).

O3. *Evaluate word embedding-based metrics for topic coherence assessment.* Conduct a comprehensive comparative study of classical co-occurrence coherence metrics and embedding-based alternatives, including static embeddings (Word2Vec, GloVe, FastText) and contextual embeddings (BERT, RoBERTa, ALBERT, MPNET). Formally evaluate all metrics by computing their correlation with ground-truth human assessments, and investigate the effect of embedding model choice, corpus selection, and transformer layer depth on coherence estimation quality.

O4. *Develop and evaluate temporal word embedding approaches for early detection of psychological disorders.* Propose a framework based on temporal word embeddings for

early risk detection that captures how users' language evolves over time. Evaluate the approach using two temporal embedding methods (TWEC and DCWE) on 18 datasets covering four mental health conditions (depression, anorexia, self-harm, and gambling). Analyze which keywords are most informative for each disorder and how their semantic contexts shift as conditions progress.

O5. *Design and implement a modular software framework for reproducible text mining experiments.* Develop LabChain, a Python framework based on pipeline architectures with automatic hash-based caching, JSON-serializable experiment configurations, and inter-team cache reuse. Validate the framework's impact on computational efficiency, reproducibility, and scientific outcomes by re-implementing and improving experiments from the research publications derived from this thesis.

2.4 Mapping of Objectives and Hypotheses to Publications

Table 2.1 provides an explicit mapping between the hypotheses, specific objectives, and the publications in which they are addressed. While each objective is primarily associated with one publication, the research program is designed so that findings and methods from earlier publications inform and support later ones.

Beyond this primary mapping, several cross-cutting relationships connect the publications:

- The clustering expertise developed under O1 directly informed the topic modeling comparison under O2, as modern topic models such as BERTopic are built on top of clustering pipelines.
- The coherence evaluation methodology developed under O3 provided the theoretical and empirical basis for the evaluation strategy employed for O2, where automatic coherence metrics are shown to correlate poorly with human judgments.
- The word embedding expertise accumulated through O2 and O3 underpinned the temporal embedding approach proposed for O4.
- The LabChain framework developed under O5 was used to re-implement the experiments developed for O4, yielding both computational savings and improved scientific results. This included the discovery and correction of a corpus preparation error that had affected prior results.

Hyp.	Obj.	Publication	Section
H1	O1	*Comparison of Clustering Algorithms for Knowledge Discovery in Social Media Publications: A Case Study of Mental Health Analysis* Procesamiento del Lenguaje Natural (SEPLN), 2024	4.1
H2	O2	*Exploiting Topic Analysis Models to Explore Psychological Dimensions in Social Media Data* Scientific Reports (Nature Portfolio), 2025	4.2
H3	O3	*A Study of Word Embedding Models for Measuring Topic Coherence* Knowledge and Information Systems (KIS), 2025 – under review	5.1
H4	O4	*Temporal Word Embeddings for Early Detection of Psychological Disorders on Social Media* Journal of Healthcare Informatics Research (JHIR), 2025	4.3
H5	O5	*LabChain: Enabling Reproducible and Modular Scientific Experiments in Python* SoftwareX (Elsevier), 2025	4.4

Table 2.1: Mapping of hypotheses and objectives to publications.

Figure 2.1 illustrates these dependencies schematically.

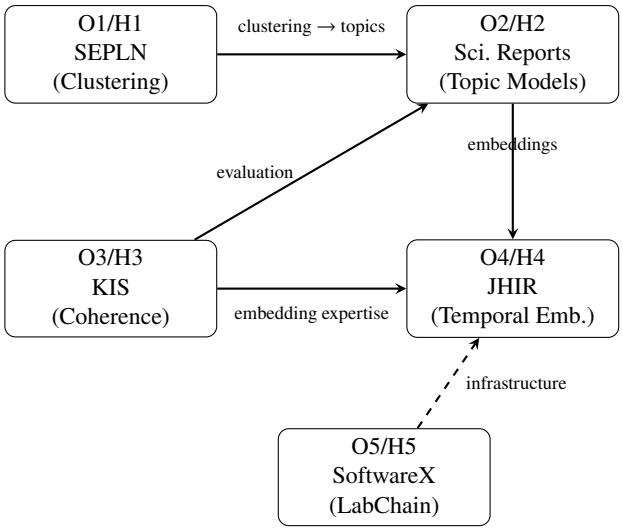


Figure 2.1: Dependencies and relationships between the research objectives and publications. Solid arrows indicate methodological flow; the dashed arrow indicates infrastructural support.

CHAPTER 3

GENERAL METHODOLOGY

This chapter describes the methodological framework shared across the five publications that constitute this thesis. Rather than repeating the specific experimental procedures of each publication –which are detailed in the respective sections– this chapter presents the common methodological elements, including data sources, text representation strategies, evaluation paradigms, and software infrastructure that supported this research project as a whole.

3.1 Research Approach

The research program follows an empirical, experimental methodology grounded in the traditions of Information Retrieval and Natural Language Processing. Each publication formulates specific research questions, proposes computational methods to address them, evaluates these methods on standardized benchmarks, and draws conclusions supported by quantitative evidence. The progression across publications follows a deliberate trajectory: from exploratory, unsupervised analysis of mental health-related text (Publications I and II) through the development of more robust evaluation tools (Publication III) to temporal modeling for early detection (Publication IV), supported by a software framework that ensures reproducibility and computational efficiency (Publication V).

A distinguishing characteristic of this research program is its emphasis on methodological rigor in evaluation. Several publications explicitly investigate the reliability of standard evaluation metrics –demonstrating, for instance, that automatic coherence metrics correlate poorly with human judgments (Publication II) and that internal clustering validation metrics cannot be meaningfully compared across different vector spaces (Publication I). This critical

stance towards evaluation pervades the entire research program and motivates the development of alternative evaluation strategies.

3.2 Data Sources

The datasets employed across the five publications can be grouped into two categories: social media corpora related to mental health, and reference corpora used for topic coherence evaluation.

3.2.1 eRisk Datasets

The primary data source throughout this thesis is the collection of datasets released by the eRisk initiative [36, 37, 47, 48, 14], a series of shared-data campaigns organized under the CLEF conference framework. These datasets comprise posts and comments collected from Reddit, where users have been identified as exhibiting signs of specific mental health conditions (positive cases) or as belonging to a control group. Four mental health conditions are covered across these research collections:

- **Depression**: Datasets from eRisk 2017, 2018, and 2022, with a total of 432 positive users and varying numbers of control users across editions.
- **Anorexia**: Datasets from eRisk 2018 and 2019, with 81 and 134 positive users, respectively.
- **Self-harm**: Datasets from eRisk 2020 and 2021, with 145 and 297 positive users, respectively.
- **Gambling**: Datasets from eRisk 2022 and 2023, with 245 and 348 positive users, respectively.

The eRisk datasets present several characteristics that make them particularly challenging for computational analysis. The class distribution is heavily imbalanced, with control users typically outnumbering positive cases by a factor of 3 to 20. Posts are written in informal English with significant variation in length, style, and topic. Users' posting histories span extended periods, enabling temporal analysis but also introducing variability and seasonal factors due to external events such as the COVID-19 pandemic.

Publications I, II, and IV exploit these datasets extensively. Publication I focuses on the eRisk 2017 depression dataset for clustering experiments. Publication II combines the eRisk 2017, 2018, and 2020 depression datasets for topic modeling, analyzing a total of 1,707 users. Publication IV evaluates temporal word embeddings across all four conditions using 18 dataset configurations spanning eRisk editions from 2017 to 2023.

3.2.2 Webis-TLDR-17 Corpus

Publication I additionally employs the Webis-TLDR-17 corpus, a large-scale collection of Reddit posts spanning multiple subreddits. A subset of 69 subreddits –containing between 5,500 and 22,000 posts each– was selected for the clustering experiments, providing a diverse sample of Reddit discourse against which mental health-related clusters could be contextualized.

3.2.3 Topic Coherence Evaluation Corpora

Publication III employs three established corpora for evaluating topic coherence metrics, each representing a different textual domain:

- **New York Times**: 47,229 news articles from May to December 2010, sourced from the GigaWord corpus. It contains 200 topics, of which 100 were evaluated by human annotators.
- **20 Newsgroups**: approximately 20,000 email messages distributed across 20 thematic categories (100 topics were generated and evaluated).
- **Genomics**: 30,000 scientific articles from the MEDLINE database, used in the TREC Genomics Track (100 topics were generated and evaluated).

Each topic is represented by 10 keywords (top words), and human quality ratings were collected from 24 to 26 annotators on a three-point scale (Useless, Average, Useful). These datasets, together with their human annotations, constitute a ground truth for assessing the correlation between automatic coherence metrics and human judgments.

3.3 Text Representation Methods

A central methodological aspect across the research program was the exploration of different text representation strategies and their impact on downstream tasks. The representations

employed span the full spectrum, from classical sparse methods to modern contextualized embeddings.

3.3.1 Classical Representations

Bag-of-words and TF-IDF representations are used as baselines in Publications I and II. These sparse, high-dimensional representations capture lexical frequency and discriminability but do not encode semantic relationships between terms. Text preprocessing for these representations follows standard NLP practice: lowercasing, removal of punctuation and numeric characters, tokenization, stopword filtering, and lemmatization using WordNet [7].

3.3.2 Static Word Embeddings

Static word embeddings –Word2Vec [42], GloVe [53], and FastText [10]– produce fixed-dimensional dense vector representations in which semantically related words occupy nearby positions in the embedding space. These models are employed in Publication III for coherence evaluation, where topic keywords are mapped to their pretrained embedding vectors and coherence is computed as a function of pairwise similarities. In Publication I, mean-pooled embeddings serve as document representations for clustering.

3.3.3 Contextualized Embeddings

Transformer-based language models –BERT [25], RoBERTa, ALBERT, and MPNET– produce context-dependent representations that capture polysemy and compositional semantics. These models are used in multiple roles across the thesis:

- As document encoders for clustering (Publication I: BERT, RoBERTa, GPT-2 with mean pooling or `[CLS]` token extraction).
- As the basis for topic coherence metrics (Publication III), where the effect of different layers and architectures on coherence estimation quality is systematically investigated.
- As the foundation for BERTopic's document representation pipeline via Sentence-BERT (Publication II).
- As the base model for dynamic contextualized word embeddings in Publication IV (DCWE, built on BERT).

3.3.4 Temporal Word Embeddings

Publication IV introduces temporal word embeddings as a representation strategy specifically designed to capture linguistic change over time. Two methods are employed:

- **TWEC** (Temporal Word Embeddings with a Compass) [26]: extends Word2Vec by training a shared "compass" embedding on the full corpus, then training time-specific input embeddings while keeping the output layer frozen. This alignment-by-construction approach enables direct comparison across time periods.
- **DCWE** (Dynamic Contextualized Word Embeddings) [34]: extends BERT with time-specific feedforward offsets applied to token representations, enabling the model to capture temporal variation in contextualized embeddings.

In both cases, a user's posting history is divided into temporal chunks, embeddings are trained for each chunk, and the displacement of keywords between temporal and reference embeddings provides features for downstream classification.

3.4 Evaluation Methodology

The evaluation strategies employed across the thesis reflect the diversity of tasks addressed and the critical attention paid to metric reliability.

3.4.1 Clustering Evaluation

Publication I employs both internal and external validation metrics:

- **Internal metrics**: Silhouette Coefficient [62], Calinski-Harabasz Index [12], and Davies-Bouldin Index [21], which assess cluster compactness and separation without reference to ground truth labels.
- **External metrics**: Adjusted Rand Index (ARI) [56] and Normalized Mutual Information (NMI) [65], which measure agreement with known labels.

As discussed in the introduction, relying on internal metrics for cross-representation comparison is problematic, since these metrics are defined relative to the distance function of the embedding space in which clustering is performed.

3.4.2 Topic Coherence Evaluation

Publications II and III address topic quality evaluation from complementary perspectives:

- **Automatic coherence metrics**: UMass, NPMI, UCI, and C_V, based on word co-occurrence statistics computed over a reference corpus or the training corpus itself. Publication II demonstrates the limitations of these metrics for social media data.
- **Embedding-based coherence metrics**: Computed as functions of pairwise embedding similarities among topic keywords, using static or contextualized embeddings. Publication III provides a systematic evaluation of these alternatives.
- **Human evaluation**: Publication II developed an annotation protocol in which six annotators (five with a Computer Science background, one with a Philology background) assessed 150 topics on a three-point Likert scale, evaluating the top words' coherence, document-topic alignment, and the overall thematic coherence. Inter-annotator agreement is measured using Cohen's Kappa.

The quality of automatic metrics is assessed by computing Spearman's rank correlation with human judgments, providing a principled basis for comparing classical and embedding-based approaches.

3.4.3 Early Risk Detection Evaluation

Publication IV evaluates temporal embedding-based detection using metrics specific to the early risk detection task:

- **F1 score**: the harmonic mean of precision and recall, providing a standard measure of classification accuracy on imbalanced data.
- **ERDE** (Early Risk Detection Error) [36]: a time-sensitive metric that applies a logistic cost function to penalize late detection of positive cases. Two configurations are used: ERDE_5 and ERDE_{50}, which control the steepness of the delay penalty.

These metrics jointly capture both the accuracy and the timeliness of risk detection, reflecting the clinical importance of early intervention.

3.4.4 Hyperparameter Optimization

Across publications, hyperparameter selection follows standard machine learning practices. Publication I uses RandomizedSearchCV with repeated stratified K-fold cross-validation. Publication IV employs grid search over classifier-specific hyperparameter spaces (e.g., regularization strength and kernel parameters for SVMs, tree depth and number of estimators for random forests) with a 70/30 train-validation split.

3.5 Software and Reproducibility

The computational experiments across the thesis rely on a consistent software ecosystem. The primary tools include:

- **scikit-learn** [51]: for clustering algorithms, classical classifiers, preprocessing utilities, and evaluation metrics.
- **Natural Language Toolkit (NLTK)** [7]: for tokenization, stopword filtering, and lemmatization.
- **Gensim** [57]: for training TWEC temporal embeddings (Word2Vec-based).
- **PyTorch** [49]: for GPU-accelerated training of DCWE temporal embeddings and neural classifiers.
- **Sentence-BERT** [58]: for document encoding in BERTopic's [31] pipeline.
- **LabChain** (Publication V): the pipeline framework developed as part of this thesis, used to re-implement and improve the experiments in Publication IV.

Publication V demonstrates that the adoption of LabChain for the temporal embedding experiments yielded both computational savings –through automatic caching of expensive intermediate computations such as BERT embeddings– and improved scientific outcomes, including the discovery and correction of a corpus preparation error that had affected prior results. This finding underscores a recurring methodological principle in this thesis: that robust software engineering practices are not merely an auxiliary concern but a direct contributor to scientific quality.

3.6 Ethical Considerations

Research involving social media data related to mental health raises important ethical questions that have been addressed throughout the research program. All experiments used pre-existing datasets released through the eRisk initiative under a signed data usage agreement. Informed consent was obtained at the time of data collection by the eRisk organisers, and not by the researchers of this thesis: participants whose Beck Depression Inventory (BDI) responses are included in the collections were recruited through a dedicated online campaign in which they were informed of the research purpose and the intended academic use of their data, voluntarily completed the questionnaire, and explicitly authorised the release of their anonymised responses together with their public social media post history for research use. The datasets were subsequently distributed to participating research teams in already anonymised form, with user identifiers removed. Accordingly, at no point did the researchers of this thesis contact users or engage with them directly; all data were accessed through the eRisk distribution channels and in compliance with the platforms' terms of service. Publication II additionally received explicit approval from the USC Ethics Committee (reference 65/2024).

The research program acknowledges several ethical limitations. The datasets are restricted to English-language content, introducing a linguistic bias that may limit the generalizability of findings. Social media users are not representative of the general population. For example, older adults, individuals without internet access, and those who use private communication channels are underrepresented. Furthermore, all computational analyses produce statistical predictions, not clinical diagnoses; the outputs of the models developed in this thesis should be interpreted as research tools that may support, but never replace, professional clinical assessment.

CHAPTER 4

RESULTS: PUBLISHED ARTICLES

This chapter presents the results published in four peer-reviewed journals. Each publication is included as a separate section: Section 4.1 presents the clustering comparison published in Procesamiento del Lenguaje Natural (SEPLN) [16]; Section 4.2 presents the topic modeling study published in Scientific Reports (Nature Portfolio) [17]; Section 4.3 presents the temporal word embedding framework published in the Journal of Healthcare Informatics Research (JHIR) [13]; and Section 4.4 presents the LabChain framework published in SoftwareX (Elsevier) [18].

4.1 Publication I: Comparison of Clustering Algorithms for Knowledge Discovery in Social Media Publications: A Case Study of Mental Health Analysis

Authors: Manuel Couto, Javier Parapar, David E. Losada
Published in: Procesamiento del Lenguaje Natural (SEPLN) [16]
Keywords: Mental Health, Social Media, Clustering, Natural Language Processing

Procesamiento del Lenguaje Natural, Revista nº 73, septiembre de 2024, pp. 69-81 recibido 08-03-2024 revisado 29-04-2024 aceptado 16-05-2024

Comparison of Clustering Algorithms for Knowledge Discovery in Social Media Publications: A Case Study of Mental Health Analysis

Comparación de algoritmos de agrupamiento para el descubrimiento de conocimiento en publicaciones de redes sociales: un caso de estudio en salud mental

Manuel Couto,[1] **Javier Parapar**[2]
David E. Losada[1]
[1]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS),
Universidade de Santiago de Compostela
[2]Centro de Investigación en Tecnoloxías da Información e da Comunicación (CITIC),
Universidade da Coruña
[1]{manuel.couto.pintos, david.losada}@usc.es
[2]javier.parapar@udc.es

Abstract: In the age of social media, user-generated content is critical for detecting early signs of mental disorders. In this study, we use thematic clustering to analyze the content of the social media platform Reddit. Our primary goal is to use clustering techniques for comprehensive topic discovery, with a focus on identifying common themes among user groups suffering from mental illnesses such as depression, anorexia, gambling addiction, and self-harm. Our findings show that certain clusters are more cohesive, e.g., with a higher proportion of texts indicating depression. Furthermore, we discovered subreddits that are strongly linked to texts from the depressed user group. These findings shed light on how online interactions and subreddit themes may impact users' mental health, paving the way for future research and more targeted interventions in the field of online mental health.
Keywords: Mental Health, Social Networks, Clustering, Natural Language Processing.

Resumen: En la era de las redes sociales, el contenido generado por los usuarios es fundamental para detectar los primeros signos de trastornos mentales. En este estudio utilizamos el agrupamiento de publicaciones por tópicos para analizar el contenido de la plataforma Reddit. Nuestro objetivo primordial es utilizar técnicas de agrupamiento para descubrir temas centrales, con un enfoque en la identificación de temas comunes entre los grupos de usuarios que sufren enfermedades mentales como la depresión, la anorexia, la adicción a los juegos de azar y las autolesiones. Nuestros hallazgos muestran que ciertos clusters son más cohesivos, por ejemplo mostrando una mayor proporción de textos de personas con depresión. Además, hemos descubierto subreddits que estan fuertemente vinculados a textos escritos por usuarios deprimidos. Estos hallazgos arrojan luz sobre cómo las interacciones en línea y los temas que se tratan en los subreddits reflejan aspectos de salud mental, abriendo el camino para futuras investigaciones e intervenciones dirigidas a la prevención de trastornos.
Palabras clave: Salud Mental, Redes Sociales, Agrupamiento, Procesamiento de Lenguaje Natural.

ISSN 1135-5948 DOI 10.26342/2024-73-5

1 Introduction

In the era of social media, the rapid growth of online platforms has provided researchers with vast amounts of user-generated data. This wealth of information presents unique opportunities for studying and understanding human behaviour, particularly in the context of psychological profiling (Crestani, Losada, and Parapar, 2022). Clustering, as a fundamental data mining technique, plays a crucial role in the analysis and organisation of such data, enabling the identification of patterns and the discovery of valuable insights.

Clustering algorithms aim to group data points into distinct clusters based on their intrinsic characteristics. These unsupervised learning methods are powerful for uncovering hidden structures within large datasets. In psychological profiling of social media publications, clustering algorithms offer the potential to identify distinct user groups, including those exhibiting signs of specific mental disorders (Aragon et al., 2021; Shensa et al., 2018; Yazdavar et al., 2017).

Commonly employed clustering algorithms can be categorised into different classes. Partition-based algorithms, including K-means and Expectation-Maximization (EM), divide the data into non-overlapping clusters based on similarity metrics (Dempster, Laird, and Rubin, 1977; MacQueen, 1967). Hierarchical algorithms, such as Agglomerative and Divisive clustering, construct cluster hierarchies through merge and split operations (Day and Edelsbrunner, 1984). Density-based algorithms, including DBSCAN and OPTICS (Ester et al., 1996; Ankerst et al., 1999), group data points based on density-connected regions, effectively handling arbitrary-shaped clusters. Model-based algorithms, such as Gaussian Mixture Models (GMM) (Reynolds, 2009), assume certain probability distributions to describe the data and infer cluster assignments.

In this study, we compare and evaluate different clustering algorithms in the context of mental health and the analysis of user publications on social media. By leveraging a taxonomy of clustering algorithms, we can systematically assess their strengths, weaknesses, and suitability for the specific task of grouping users' posts, with a particular focus on texts related to various mental disorders.

The results of our study show that certain clusters share a strong thematic homogeneity (particularly topics focusing on romantic relationships or video games). On the other hand, we have seen that certain clusters are composed of a larger proportion of texts posted by depressed users.

In addition, we have been able to understand which communities (subreddits) are most correlated with the group of depressed users. This provides useful insights on which communities might have the largest proportion of depressed users or what kind of problems depressed people might be dealing with.

2 Related Work

The analysis of user-generated content on social media offers unique opportunities for psychological profiling, as it provides rich and extensive data about individuals' thoughts, emotions, and behaviour in an online environment (Crestani, Losada, and Parapar, 2022; Chancellor and De Choudhury, 2020; Parapar et al., 2022; Couto, Pérez, and Parapar, 2022). By leveraging clustering algorithms, researchers can identify distinct user groups and gain insights into the psychological characteristics of individuals (Clatworthy et al., 2005). However, conducting psychological profiling using social media data poses challenges due to the large volume of unstructured text and the need for accurate representation and analysis of user posts. The selection of the appropriate clustering algorithm depends on the nature of the data and the objectives of the study. In general-purpose clustering applications, partition-based algorithms are commonly used for their simplicity and efficiency, while density-based algorithms are effective in handling datasets with irregularly shaped clusters. Model-based algorithms assume specific probability distributions, making them suitable for certain applications such as topic modelling in psychological profiling (Fahad et al., 2014; Ezugwu et al., 2022).

Clustering algorithms have shown promise in identifying user groups exhibiting symptoms of various mental disorders, including depression, anxiety, eating disorders, addiction, and self-harm (Nguyen et al., 2022; Aragón, López-Monroy, and Montes-y Gómez, 2019; Aragon et al., 2021; Peres et al., 2021; Ghaharian et al., 2022; Crestani, Losada, and Parapar, 2022; Ríssola, Losada, and Crestani, 2021). By clustering users ba-

sed on their textual content, these studies have revealed distinct patterns and characteristics associated with different mental health concerns.

Evaluating the quality and effectiveness of clustering algorithms in psychological profiling studies requires appropriate evaluation metrics and validation techniques. Commonly used metrics include internal validation metrics such as Silhouette coefficient (Rousseeuw, 1987), Calinski-Harabasz Index (Caliński and Harabasz, 1974) or Davies-Boulding Index (Davies and Bouldin, 1979), which assess the compactness and separation of clusters. As external validation metrics, the Rand index (Rand, 1971), Normalize Mutual Information (Strehl and Ghosh, 2002) and Purity (Marutho et al., 2018) are commonly utilised to quantify the alignment between clustering outcomes and ground truth labels, thus capturing the agreement between them. Additionally, other validation techniques such as external indices (e.g., F-measure), internal indices (e.g., Dunn index), and visual inspection of cluster visualisations are employed to evaluate clustering solutions (Fahad et al., 2014; Emmons et al., 2016; Palacio-Niño and Berzal, 2019).

Studies applying clustering algorithms for psychological profiling on social media employ these metrics to assess the accuracy, reliability, and transferability of the obtained clustering results. For instance, Gao et al. (2023) provides a comprehensive review of methods and guidelines for employing clustering algorithms in the context of mental health. The authors emphasise the importance of internal and external validation metrics. Another example is the work by Ikeda et al. (2013), where hybrid methods are employed for user profiling on Twitter. In this study, clustering techniques are used to generate user groupings based on follower-followee relationships. Precision, recall, and F-measure were utilised as external validation metrics to evaluate the performance of the system. By employing appropriate evaluation measures, researchers can determine the effectiveness and applicability of different clustering algorithms in user profiling.

The present study aims to contribute to the advancement of psychological profiling techniques on social media platforms and improve the identification and support of individuals with mental health concerns. While other works focused on solving a classification problem, in this work we study the forms of representation and algorithms that behave best, with the aim of conducting an exploratory study. The primary objective of this study is to extract valuable insights from data. This knowledge enables us to gain a better understanding of the topics discussed by Reddit users experiencing depression. We explore the intricate correlations between Reddit communities and users displaying signs of depression. These findings are informative about mental health discussions in online communities and provide a valuable foundation for the development of effective tools for early detection of mental health issues.

3 Methodology

To obtain a diverse set of social media contents, we used two different **data collections**, the Webis-TLDR-17 collection (Völske et al., 2017) and the eRisk 2017 depression collection (Losada, Crestani, and Parapar, 2017).

Webis-TLDR-17. This collection consists of multiple posts from the social network Reddit, where each post has information about the subreddit (Reddit's subcommunity) where it was posted and a summary of the publication.

eRisk 2017 depression. This collection contains a thread of posts or comments from multiple Reddit users. Each user is annotated with a label indicating whether or not the user was diagnosed with depression (positive group or control group, respectively).

In this paper we thus work with texts that can be either posts or comments. We will use the generic term text or publication to refer to any individual post or individual comment. Any text is always linked to a user and a class. In the case of the Webis-TLDR-17 dataset, the class refers to the subreddit where the text was published. In the case of the eRisk 2017 depression dataset, the class refers to whether the text comes from a depressed user or a control user. We performed **sampling** on the Webis-TLDR-17 dataset as it contains millions of texts. We grouped all the content of the collection by subreddits and retained only those subreddits that had between 5,500 and 22,000 texts. This sampling allowed us to reduce the size of the dataset and focus on representative subreddits that have a substantial number of texts. By selec-

ting this specific range, we aimed to balance the availability of texts for each subreddit, avoiding subreddits with too few or too many entries. As a result, we ended up with 69 different subreddits.

The **preprocessing** stage plays a crucial role in cleaning and transforming the raw text, ensuring that it is suitable for subsequent analysis. We removed special characters with regular expressions, tokenised the texts into individual words or n-grams, eliminated stopwords, and normalised the text through techniques such as lemmatisation or stemming with the NLTK library, version 3.8.1 (Bird, Klein, and Loper, 2009).

Once the text data was cleaned, we employed various **vectorisation** techniques to represent the texts as numerical vectors suitable for clustering algorithms. We explored different alternatives, including:

Term Frequency (TF): The TF approach represents each document as a vector, with each dimension corresponding to a unique word in the corpus. Each numerical value represents the frequency of that word within the document. The final dimensionality of the TF vectors is equal to the vocabulary size (Croft, Metzler, and Strohman, 2010).

Term Frequency-Inverse Document Frequency (TF-IDF): The TF-IDF technique multiplies the term frequency by the inverse document frequency to represent the importance of words in a document. It assigns higher weights to words that are frequent within a document but rare across the entire corpus. The vector dimensionality is the same as that of the TF vectors (Croft, Metzler, and Strohman, 2010).

Bidirectional Encoder Representations from Transformers (BERT): It is a pre-trained language model that generates contextualised word embeddings. It captures the contextual meaning of words by considering their surrounding context within a sentence or document. BERT-based embeddings have been widely used in various natural language processing tasks, including clustering. The dimensionality of the $\text{BERT}_{\text{base}}$ embeddings used for this work is 768 dimensions (Devlin et al., 2018).

RoBERTa, which is another pre-trained language model that extends BERT's architecture. It incorporates additional pre-training techniques and achieves improved performance on various language understanding tasks. RoBERTa-based embeddings capture more nuanced contextual information and can enhance clustering results. The dimensionality of the $\text{RoBERTa}_{\text{base}}$ embeddings used for this work is 768 dimensions (Liu et al., 2019).

Generative Pre-trained Transformer 2 (GPT-2): GPT-2 is a powerful language model that generates coherent and contextually relevant text. It can be used to generate word embeddings that capture the semantic meaning of words and sentences. GPT-2-based embeddings have shown promising results in clustering tasks. The dimensionality of the GPT-2 embeddings used for this work is 768 dimensions (Radford et al., 2019).

TF and TF-IDF allow representing texts as vectors in a classic manner without any further computation step. For BERT and RoBERTa, we utilised mean pooling to create sentence embeddings (Devlin et al., 2018; Liu et al., 2019). Mean pooling calculates the average of all the word embeddings in a sentence, resulting in a fixed-length representation that captures the overall meaning of the sentence. This approach allows us to generate sentence embeddings that encapsulate the contextual information derived from the entire sequence. For GPT-2, we used the special token [CLS] as a summary of the sequence (Radford et al., 2019). The [CLS] token represents the entire sequence and carries information about the context and meaning of the text. By extracting the embedding of the [CLS] token, we obtain a condensed representation that captures the salient features and overall semantic meaning of the sequence.

By employing these vectorisation techniques (TF, TF-IDF, BERT, RoBERTa, and GPT-2), we aimed to capture different aspects of the textual data, including frequency-based importance, contextual understanding, and semantic meaning. These representations enable a more comprehensive analysis of the problem and facilitate effective clustering of the user-generated content.

In order to enhance the efficiency and effectiveness of our methodology, we incorporated the possibility of performing a **dimensionality reduction** step after vectorisation. This step aims to reduce the dimensionality of the text representations while preserving the most informative features.

We employed Singular Value Decomposition (SVD) as the chosen technique for di-

mensionality reduction. SVD decomposes the matrix of vectorised texts into three matrices representing the singular values, left singular vectors, and right singular vectors. By selecting a subset of the top-k singular values and corresponding singular vectors, we effectively reduced the dimensionality of the vectorised data.

The application of dimensionality reduction using SVD serves two primary purposes. Firstly, it significantly reduces the computational complexity of subsequent clustering algorithms, facilitating faster experimentation. Secondly, it can improve the performance of the clustering algorithms by eliminating noisy and irrelevant features, leading to more accurate and meaningful clusters (Ding and He, 2004; Kadhim, Cheah, and Ahamed, 2014).

3.1 Clustering Algorithms

In this section, we discuss the clustering algorithms employed to group the texts extracted from Reddit into meaningful clusters. Clustering is an unsupervised learning technique that aims to discover inherent patterns and structures in the data, allowing us to identify similarities and differences between the documents.

We explore several popular clustering algorithms known for their effectiveness in text analysis tasks. Each algorithm employs a unique approach to partition the data points into clusters based on their similarity. The choice of clustering algorithm depends on various factors, including the dataset characteristics, scalability, interpretability, and the desired clustering outcomes (Fahad et al., 2014; Ezugwu et al., 2022; Mahdi, Hosny, and Elhenawy, 2021).

The following clustering algorithms were selected for this study:

K-means is a widely used centroid-based clustering algorithm that aims to partition the data into a predefined number of clusters (K). It iteratively assigns data points to the nearest cluster centroid and updates the centroids until convergence. K-means is known for its simplicity, efficiency, and effectiveness in finding spherical-shaped clusters (Arthur and Vassilvitskii, 2006).

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a density-based clustering algorithm that groups data points based on their density in the feature space. It identifies dense regions as clusters and considers low-density regions as noise or outliers. DBSCAN is effective in discovering clusters of arbitrary shapes and handling datasets with varying densities (Ester et al., 1996).

Gaussian Mixture Models (GMM) assume that the data points are generated from a mixture of Gaussian distributions. The GMM clustering algorithm fits a specified number of Gaussian components to the data, assigning data points to the most probable cluster. GMM is versatile, capturing clusters with different shapes and accommodating overlapping clusters (Reynolds, 2009).

By employing these diverse clustering algorithms, we aim to explore different approaches to organise the text data into clusters. Each algorithm brings unique characteristics and capabilities, enabling us to uncover distinct structures. While other algorithms were considered, such as Spectral clustering (Bolla, 2013) and Affinity propagation (Frey and Dueck, 2007), Agglomerative (Nielsen and Nielsen, 2016; Murtagh and Contreras, 2012), or Birch (Zhang, Ramakrishnan, and Livny, 1996), they were not included in this study due to their computational and memory limitations when working with large datasets (Mahdi, Hosny, and Elhenawy, 2021; Fahad et al., 2014).

3.2 Evaluation metrics

Internal validity metrics focus on evaluating the quality and coherence of the clusters based on the intrinsic characteristics of the data. The following internal validity metrics were utilised in our evaluation:

Silhouette Coefficient: The Silhouette coefficient (Rousseeuw, 1987) measures the average similarity of each data point to its own cluster compared to other clusters. It ranges from -1 to 1, with higher values indicating better-defined and more cohesive clusters.

Calinski-Harabasz Index (CH): The CH index (Caliński and Harabasz, 1974) computes the ratio of between-cluster dispersion to within-cluster dispersion. Higher values of this index indicate better-defined clusters with greater separation.

Davies-Bouldin Index (DB): DB index (Davies and Bouldin, 1979) measures the average similarity between clusters and provides a measure of the cluster separation. Lower

values indicate better-defined and more separated clusters.

External validity metrics assess the clustering results in relation to a ground truth with known class labels. These metrics measure the agreement between the clustering assignments and the true labels. The following external validity metrics were employed in our evaluation:

Adjusted Rand Index (ARI): The ARI (Hubert and Arabie, 1985) quantifies the similarity between the clustering assignments and the true labels. It considers all pairs of samples and evaluates the agreement between them, providing a value between -1 and 1. A higher value indicates better agreement.

Normalized Mutual Information (NMI): The NMI (Strehl and Ghosh, 2002) measures the amount of mutual information shared between the clustering assignments and the true labels, taking into account the class distribution. It ranges from 0 to 1, with higher values indicating better agreement.

3.3 Experimentation setup

The experimentation stage was divided into two phases. In the first phase, our objective was to determine the best combinations of vectorisers, clustering algorithms, and hyperparameters. Therefore, the space for experimentation was vast. We have optimized the number of clusters in all the models (testing the range between 2 and 10, in steps of 1, and the range between 10 and 100, in steps of 10). In addition, we tested two tolerance parameter configurations (values of 10^{-3} and 10^{-2}) in both GMM and k-means. We have also optimized certain model-specific parameters, such as the hyperparameters "algorithm" in kmeans, the "covariance type" in gmm, or the "eps" and "neightbours" of DBSCAN. However, performing a full exploration of hyperparameters would require a significant amount of time. To address this challenge, we adopted a non-exhaustive search approach to find optimal hyperparameter configurations. We utilized the RepeatedStratifiedKFold[1] algorithm and RandomizedSearchCV[2] technique, guided by the external validation metrics ARI and NMI. This allowed us to efficiently explore the hyperparameter space and identify promising configurations without exhaustively evaluating every possible combination. The search was performed on the Webis-TLDR-17 dataset, where the labels correspond to the subreddits from which the texts were extracted. Once we identified the best clustering configuration based on the validation metrics, we proceeded to a second phase, in which we leveraged the optimal clustering results obtained from the previous phase to conduct an exploratory analysis of the data.

[1] `https://scikit-learn.org/stable/modules/cross_validation.html`.

[2] `https://scikit-learn.org/stable/modules/grid_search.html`.

4 Clustering results

We begin by providing a summary of the best combination of clustering algorithm and vectorial representation. Table 1 reports the performance metrics obtained for the best configurations. It can be noted that the vectorial representation that has achieved the best results in terms of internal validation is TF. This could be attributed to a potential bias towards text length, as the TF approach emphasizes the importance of individual terms within the documents without any kind of normalisation. The TF feature values grow with no bound, while TF-IDF, BERT, RoBERTa, and GPT-2 have their vectorial representation values normalised (because of the TF-IDF weighting scheme or the layer normalisation technique in the transformer architecture).

On the other hand, when considering external validation metrics, it is observed that RoBERTa performs the best among the vectorial representations. RoBERTa, being a transformer-based language model, is capable of capturing more nuanced semantic information, which likely contributes to its superior performance in capturing the underlying patterns in the data.

Additionally, in terms of external validation, it is generally observed that GMM and K-means clustering outperform DBSCAN. This suggests that GMM and K-means are more effective in capturing the structure and patterns of the user-generated content compared to DBSCAN.

After conducting a non-exhaustive search for the best configurations, we have found that the combination yielding the highest ARI and NMI score is the GMM model with RoBERTa's representation. The optimal hyperparameters for this configuration were as

		External		Internal		
Algorithm	Vectorisation	ARI	NMI	Silhouette	CH	DB
GMM	TF N:3	0.096	0.092	0.430	236614.452	0.806
K-means	TF N:3	0.053	0.087	0.392	**258067.923**	0.8671
DBSCAN	TF N:3	0.008	0.002	**0.631**	233440.350	**0.579**
GMM	TF-IDF N:3	0.151	0.172	0.281	118091.257	1.129
K-means	TF-IDF N:3	0.126	0.164	0.311	132990.815	1.049
DBSCAN	TF-IDF N:3	-0.009	0.025	0.092	9685.614	2.341
GMM	BERT N:3	0.094	0.123	0.277	126807.031	1.072
K-means	BERT N:3	0.102	0.108	0.361	148423.361	0.965
DBSCAN	BERT N:3	0.001	0.001	-0.153	219.810	117.855
GMM	RoBERTa N:3	**0.180**	**0.200**	0.287	141981.974	1.211
K-means	RoBERTa N:3	0.169	0.183	0.293	149548.197	1.104
DBSCAN	RoBERTa N:3	0.078	0.147	0.200	99696.919	1.513
GMM	GPT-2 N:3	0.010	0.015	0.252	156534.401	1.129
K-means	GPT-2 N:3	0.010	0.015	0.256	226903.435	0.983
DBSCAN	GPT-2 N:3	-0.001	0.000	-0.153	187.103	336.351

Tabla 1: Best Hyperparameter Configurations and Performance Metrics (Webis-TLDR-17).

follows: Covariance type: "tied", Initialization parameters: "kmeans", Number of components: 5, Tolerance: 0.001.

Next, we proceed with an analysis phase of this clustering approach. We clustered the entire Webis-TLDR-17 corpus using this configuration and we counted the number of texts from each subreddit. By visualising this information for each cluster, we can gain insights into the distribution of subreddits across the identified clusters. By examining these trends, we can identify dominant or underrepresented subreddits within specific clusters, which could indicate shared themes or topics among the groups.

In order to accurately visualise the distribution of subreddits across different clusters, we have generated bar graphs or histograms. These histograms, shown in Figure 1, represent the most prominent subreddits in a every cluster. This visualisation provides interesting insights. For instance, it is evident that Cluster 1 and Cluster 4 are clusters with a predominant theme.

Cluster 1 is mainly composed of texts from subreddits such as relationships, relationship_advice, sex, and dating_advice, indicating a clear thematic homogeneity. Note also that the relationships subreddit has also a strong presence in clusters 2 and 3 but these other clusters show a more varied set of subreddits.

Cluster 4 shows a prominent subreddit leagueoflegends and other relevant subreddits related to various video games. This cluster has a clear thematic focus on video gaming. The leagueoflegends subreddit and other video game-related subreddits are also significantly present in other clusters, specifically clusters 0 and 3.

Clusters 0, 2, and 3 consist of subreddits covering a wider range of topics. For instance, we can see that cluster 0 includes texts from subreddits such as explainlikeimfive, atheism, Fitness, askscience, leagueoflegends, DnD, trees and so forth. These subreddits are not associated to a single theme but, instead, represent communities oriented to asking questions and engaging in discussions about various subjects. In the case of Cluster 2, we can see prominent subreddits such as Personalfinance, politics, relationships, explainlikeimfive, adviceanimals, worldnews or legaladvice. This group does not represent a focused set of discussions but it seems to include more formal subjects such as politics, finance, and even divorce-related issues. Lastly, Cluster 3 is predominantly composed of subreddits such as tifu, relationships, trees, leagueoflegends, fitness and adviceanimals. Once again, there is no specific common topic, and this group includes texts related to complaints or grievances about relationships, plant care, animal-related discussions, or following a training plan.

While certain clusters exhibit a strong thematic coherence, other clusters comprise texts from highly different sources. This suggests the existence of cross-discussions or shared interests among different clusters, highlighting the diverse nature of user-generated content within the Webis-TLDR-17 dataset.

Additionally, we analysed the pattern of assignment of texts from 69 subreddits to the five clusters. This analysis, presented in Figure 2, allows us to understand, for example, whether a given subreddit is concentrated on a single cluster or, instead, dispersed

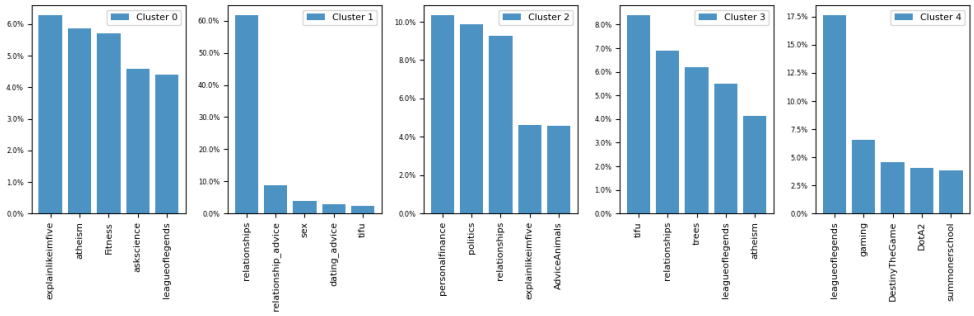


Figura 1: The subreddits with the highest proportion of texts in each cluster.

across several clusters. By understanding the distribution of each subreddit's texts within clusters, we gain insights into the subreddit's content cohesion and heterogeneity.

We can observe that certain subreddits such as Adviceanimals, IAmA, changemyview, or Worldnews are dispersed across multiple clusters. On the other hand, subreddits like DestinyTheGame, askscience, dating_advice, relationships and personalfinance are highly concentrated in specific clusters.

Some interesting observations can be made from the correlation matrix. For instance, there are high correlation values between the subreddit electronic_cigarette and various video game-related subreddits. This could be indicative of a potential overlap or shared interests among individuals who engage in both activities. There could also exist a subset of users who are active participants in discussions related to both electronic cigarettes and video games. Additionally, there is a notable correlation between the subreddit for depression and subreddits such as AskMen, AskWomen, TwoXChromosomes, self and sex. This suggests a connection between mental health discussions and topics related to relationships, gender, and self-expression. In fact, individuals seeking support or information about depression can be prone to engage in conversations about these related subjects.

4.1 eRisk collection

The clustering organisation described above was further exploited to analyse the publications from the eRisk 2017 depression dataset. To that end, we report here about the assignment of writings posted by different categories of eRisk users (depressed vs non-depressed) into the previously obtained clusters. By imputing each eRisk text to its closest cluster we can gain further insights into the posting patterns of users experiencing depression and try to understand how this reflects on different Reddit communities. This imputation process enables us to associate the depression-related texts with specific clusters, facilitating a comprehensive analysis of the data. The results of the imputation can be observed in Figure 3. Clusters 1 and 3 exhibit a higher proportion of texts from the depressed group compared to the control group. Note that these two clusters (see Figure 1) have a substantial portion of texts discussing personal experiences (e.g., relationships).

Next, we focus on the comparison between the depressed group and different subreddits. Figure 4 shows the correlation between the depressed group and several subreddits. These correlations were estimated by comparing the distribution across clusters of the depressed publications and the distribution across clusters of the subreddit's publications. Then we ordered in descending order the correlations. Notably, subreddits such as trees, NoFap, Drugs, ADHD, tifu, and talesfromtechsupport yield stronger correlations with the depressed group.

5 Discussion and Conclusions

After conducting this exploratory study, we have obtained relevant information from both datasets. Firstly, from the perspective of the formed clusters, we gained insights into the thematic content of each cluster. Secondly, from the perspective of the subreddits, we acquired statistical information about each subreddit, enabling us to compare them based on the distributions of their texts across dif-

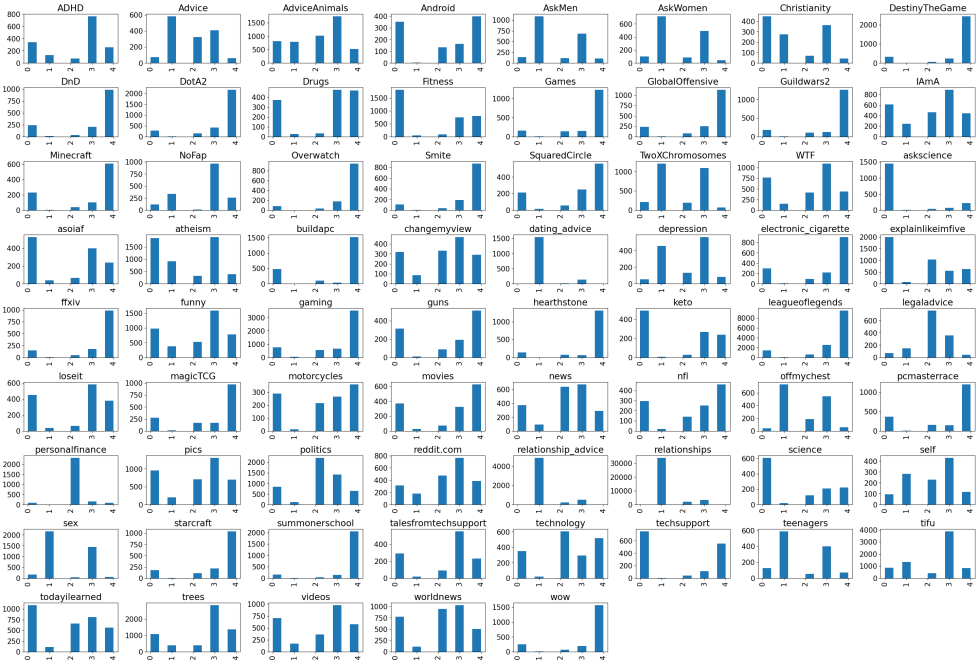


Figura 2: Histograms representing the distribution of the subreddit's texts in the clusters.

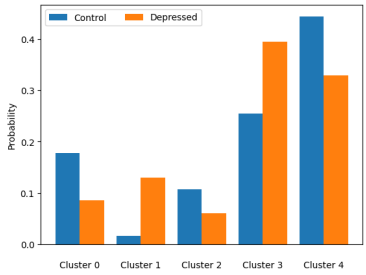


Figura 3: Distribution of probabilities of the texts from a certain user group (depressed/control) across the clusters.

ferent clusters.

As a result of this analysis, several significant findings have emerged. Clusters 1 and 4 are somehow focused clusters that share texts from various subreddits, focusing on the themes of relationships and video games, respectively.

Furthermore, after imputing the depression texts into the previously constructed clusters, we have obtained additional relevant information. From the perspective of the clusters, we observed that clusters 1 and 3 have a higher proportion of texts from depressed users compared to control users. This finding could potentially help to detect at-risk users. For instance, the activity of users within certain communities –not necessarily related to mental health– could be informative and, thus, act as supporting evidence or define new predictive features.

Texts from the depressed group exhibit a strong correlation with subreddits like trees, NoFap, Drugs, and ADHD. Some relevant words from these subreddits are shown in Figure 5 in the form of a wordcloud. Interestingly, the subreddit yielding the highest correlation is trees which, at first glance, may not seem directly related to depression. This subreddit is mainly focused on cannabis consumption. Smoking marijuana is closely associated with a variety of mental diseases, including depression. The Drugs subreddit is a similar case. Drug use is the primary topic of discussion in this forum; but substance abuse has been often linked to the development of mental disorders. The other subreddits are slightly different. NoFap is a pornography and sex addiction forum. On the other hand, ADHD, is a community where

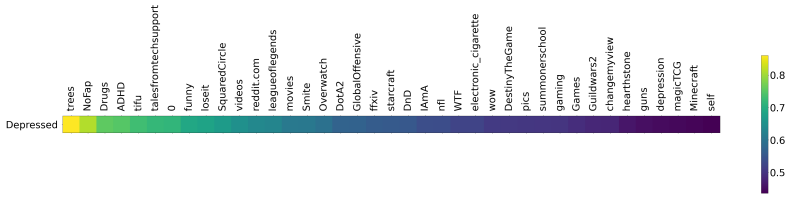
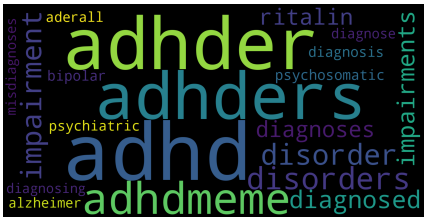
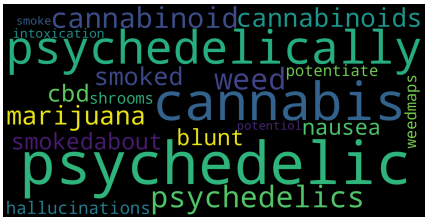


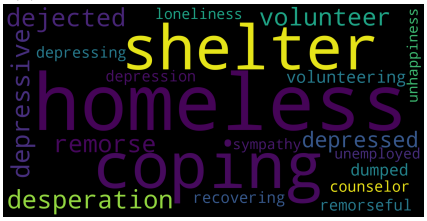
Figura 4: Most Correlated subreddits to the depressed set.



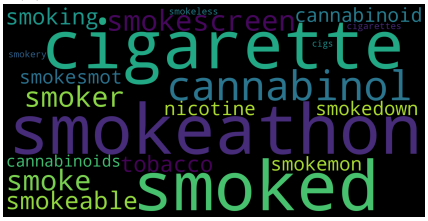
(a) Word cloud from the ADHD subreddit.



(b) Word cloud from the Drugs subreddit.



(c) Word cloud from the NoFap subreddit.



(d) Word cloud from the Trees subreddit.

Figura 5: Word cloud of the main topics from the most relevant subreddits.

people with ADHD can share their stories, struggles, and non-medication solutions. All of these forums are about problems, addictions, and people sharing negative experiences, life struggles, or something about themselves that makes them unhappy.

Acknowledgements

This work was supported by project PLEC2021-007662 (MCIN/AEI/10.13039/501100011033, Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, Plan de Recuperación, Transformación y Resiliencia, Unión Europea-Next Generation EU). The first and third author thank the financial support supplied by the Xunta de Galicia-Consellería de Cultura, Educación, Formación Profesional e Universidade (ED431G 2023/04, ED431C 2022/19) and the European Regional Development Fund, which acknowledges the CiTIUS-Research Center in Intelligent Technologies of the University of Santiago de Compostela as a Research Center of the Galician University System. David E. Losada also thanks the financial support obtained from project SUBV23/00002 (Ministerio de Consumo, Subdirección General de Regulación del Juego) and project PID2022-137061OB-C22 (Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, Proyectos de Generación de Conocimiento; supported by the European Regional Development Fund).

The second author thanks the financial support supplied by the the CITIC, as a center accredited for excellence within the Galician University System and a member of the CIGUS Network, funded by the Xunta de Galicia and the EU through the ERDF Galicia 2021-27 operational program (ref. ED431G 2023/01) and project PID2022-137061OB-C21 (MCIN/AEI/ 10.13039/501100011033, Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, Proyectos de Generación de Conocimiento; supported by “ERDF A way of making Europe”, by the “European Union”).

References

Ankerst, M., M. M. Breunig, H.-P. Kriegel, and J. Sander. 1999. Optics: Ordering points to identify the clustering structure. *ACM Sigmod record*, 28(2):49–60.

Aragon, M. E., A. P. Lopez-Monroy, L.-C. G. Gonzalez-Gurrola, and M. Montes. 2021. Detecting mental disorders in social media through emotional patterns-the case of anorexia and depression. *IEEE Transactions on Affective Computing*.

Aragón, M. E., A. P. López-Monroy, and M. Montes-y Gómez. 2019. Inaoe-cimat at erisk 2019: Detecting signs of anorexia using fine-grained emotions. In *CLEF (Working Notes)*.

Arthur, D. and S. Vassilvitskii. 2006. k-means++: The advantages of careful seeding. Technical report, Stanford.

Bird, S., E. Klein, and E. Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Bolla, M. 2013. *Spectral clustering and biclustering: Learning large graphs and contingency tables*. John Wiley & Sons.

Caliński, T. and J. Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27.

Chancellor, S. and M. De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, 3(1):43.

Clatworthy, J., D. Buick, M. Hankins, J. Weinman, and R. Horne. 2005. The use and reporting of cluster analysis in health psychology: A review. *British journal of health psychology*, 10(3):329–358.

Couto, M., A. Pérez, and J. Parapar. 2022. Temporal word embeddings for early detection of signs of depression. In *Proceedings of the CIRCLE (Joint Conference of The Information Retrieval Communities in Europe)*.

Crestani, F., D. E. Losada, and J. Parapar. 2022. *Early Detection of Mental Health Disorders by Social Media Monitoring: The First Five Years of the ERisk Project*, volume 1018. Springer Nature.

Croft, W. B., D. Metzler, and T. Strohman. 2010. *Search engines: Information retrieval in practice*, volume 520. Addison-Wesley Reading.

Davies, D. L. and D. W. Bouldin. 1979. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227.

Day, W. H. and H. Edelsbrunner. 1984. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification*, 1(1):7–24.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ding, C. and X. He. 2004. K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on Machine learning*, page 29.

Emmons, S., S. Kobourov, M. Gallant, and K. Börner. 2016. Analysis of network clustering algorithms and cluster quality metrics at scale. *PloS one*, 11(7):e0159161.

Ester, M., H.-P. Kriegel, J. Sander, X. Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231.

Ezugwu, A. E., A. M. Ikotun, O. O. Oyelade, L. Abualigah, J. O. Agushaka, C. I. Eke, and A. A. Akinyelu. 2022. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, 110:104743.

Fahad, A., N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A. Y. Zomaya, S. Foufou, and A. Bouras. 2014. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE transactions on emerging topics in computing*, 2(3):267–279.

Frey, B. J. and D. Dueck. 2007. Clustering by passing messages between data points. *science*, 315(5814):972–976.

Gao, C. X., D. Dwyer, Y. Zhu, C. L. Smith, L. Du, K. M. Filia, J. Bayer, J. M. Menssink, T. Wang, C. Bergmeir, et al. 2023. An overview of clustering methods with guidelines for application in mental health research. *Psychiatry Research*, page 115265.

Ghaharian, K., B. Abarbanel, D. Phung, P. Puranik, S. Kraus, A. Feldman, and B. Bernhard. 2022. Applications of data science for responsible gambling: a scoping review. *International Gambling Studies*, pages 1–24.

Hubert, L. and P. Arabie. 1985. Comparing partitions. *Journal of classification*, 2:193–218.

Ikeda, K., G. Hattori, C. Ono, H. Asoh, and T. Higashino. 2013. Twitter user profiling based on text and community mining for market analysis. *Knowledge-Based Systems*, 51:35–47.

Kadhim, A. I., Y.-N. Cheah, and N. H. Ahamed. 2014. Text document preprocessing and dimension reduction techniques for text document clustering. In *2014 4th international conference on artificial intelligence with applications in engineering and technology*, pages 69–73. IEEE.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Losada, D. E., F. Crestani, and J. Parapar. 2017. erisk 2017: Clef lab on early risk prediction on the internet: experimental foundations. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11–14, 2017, Proceedings 8*, pages 346–360. Springer.

MacQueen, J. 1967. Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pages 281–297. University of California Los Angeles LA USA.

Mahdi, M. A., K. M. Hosny, and I. Elhenawy. 2021. Scalable clustering algorithms for big data: A review. *IEEE Access*, 9:80015–80027.

Marutho, D., S. H. Handaka, E. Wijaya, et al. 2018. The determination of cluster number at k-mean using elbow method and purity evaluation on headline news. In *2018 international seminar on application for technology of information and communication*, pages 533–538. IEEE.

Murtagh, F. and P. Contreras. 2012. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1):86–97.

Nguyen, T., A. Yates, A. Zirikly, B. Desmet, and A. Cohan. 2022. Improving the generalizability of depression detection by leveraging clinical questionnaires. *arXiv preprint arXiv:2204.10432*.

Nielsen, F. and F. Nielsen. 2016. Hierarchical clustering. *Introduction to HPC with MPI for Data Science*, pages 195–211.

Palacio-Niño, J.-O. and F. Berzal. 2019. Evaluation metrics for unsupervised learning algorithms. *arXiv preprint arXiv:1905.05667*.

Parapar, J., P. Martín-Rodilla, D. E. Losada, and F. Crestani. 2022. erisk 2022: pathological gambling, depression, and eating disorder challenges. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part II*, pages 436–442. Springer.

Peres, F., E. Fallacara, L. Manzoni, M. Castelli, A. Popovič, M. Rodrigues, and P. Estevens. 2021. Time series clustering of online gambling activities for addicted users' detection. *Applied Sciences*, 11(5):2397.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rand, W. M. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

Reynolds, D. A. 2009. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663).

Ríssola, E. A., D. E. Losada, and F. Crestani. 2021. A survey of computational methods for online mental state assessment on social media. *ACM Trans. Comput. Healthcare*, 2(2), mar.

Rousseeuw, P. J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Shensa, A., J. E. Sidani, M. A. Dew, C. G. Escobar-Viera, and B. A. Primack. 2018. Social media use and depression and anxiety symptoms: A cluster analysis. *American journal of health behavior*, 42(2):116–128.

Strehl, A. and J. Ghosh. 2002. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617.

Völske, M., M. Potthast, S. Syed, and B. Stein. 2017. Tl; dr: Mining reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63.

Yazdavar, A. H., H. S. Al-Olimat, M. Ebrahimi, G. Bajaj, T. Banerjee, K. Thirunarayan, J. Pathak, and A. Sheth. 2017. Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*, pages 1191–1198.

Zhang, T., R. Ramakrishnan, and M. Livny. 1996. Birch: an efficient data clustering method for very large databases. *ACM sigmod record*, 25(2):103–114.

4.2 Publication II: Exploiting Topic Analysis Models to Explore Psychological Dimensions in Social Media Data

Authors: Manuel Couto, Javier Parapar, David E. Losada
Published in: Scientific Reports (Nature Portfolio) [17]
Keywords: Social Media, Mental Health, Psychological Analysis, Topic Analysis, Large Language Models

scientific reports

Check for updates

OPEN

Exploiting topic analysis models to explore psychological dimensions in social media data

Manuel Couto[1]✉, Javier Parapar[2] & David E. Losada[1]

Automatic topic generation is a fundamental tool in unstructured text analysis, yet its application to noisy web-based collections for extracting psychological patterns remains underexplored. This work compares three representative topic models from different families: Latent Dirichlet Allocation (classical probabilistic), BERTopic (embedding-based), and TopClus (deep neural network), evaluating their performance on mental health data from the eRisk initiative. Using posts from individuals with depressive disorders and control groups, we assess topic quality through both automatic coherence metrics and rigorous human evaluation by expert reviewers. This dual approach addresses the limitations of purely automatic evaluation in complex social media datasets where thematic content does not always reveal psychological cues. Our results demonstrate that BERTopic significantly outperforms other models in perceived coherence, identifying clearer and more specific themes, including depression-related topics such as mental health struggles and self-harm. Thematic analysis across user groups revealed that certain topics contained higher proportions of posts from individuals with depression, providing actionable insights for psychological screening. This work underscores the potential of advanced topic models for mental health analysis in noisy social media data and highlights the importance of human evaluation in validating topic quality for sensitive applications.

Keywords Social media, Mental health, Psychological analysis, Topic analisis, Large language models

Automatic topic generation is a fundamental tool for analyzing unstructured text, enabling researchers to uncover patterns and underlying themes in large datasets. The generation of high-quality topics is a complex task within Natural Language Processing (NLP) and it has gained significant momentum in recent years following the development of advanced neural models[1–5]. Within the mental health domain, the ability to identify latent themes expressed in user-generated content is particularly relevant, as it can reveal people's concerns, coping strategies, or symptomatology that are not always explicit in individual posts. These insights may complement clinical assessments, support early risk detection, and provide valuable resources for mental health professionals and social scientists[6–10]. Given the high volume of posts on social media, manual monitoring is impractical, underscoring the need for automated technologies to monitor recurring problems and their evolution. Despite its potential, research on topic modeling specifically applied to mental health remains limited[11–13]. There is a need for systematic evaluations of topic modeling methods in challenging and noisy settings such as social media. In these types of online environments, relevant signals are often subtle and obscured by informal language.

The challenges in this area lie not only in designing models capable of identifying relevant thematic patterns, but also in adequately evaluating the quality of the generated topics. While automatic metrics are useful and less costly than human judgments, their limited reliability in capturing semantic coherence has been well documented[14,15]. These metrics become particularly problematic when evaluating topics generated from noisy corpora –such as texts from social networks– that frequently contain slang and spelling errors.

Different approaches have been proposed to address the task of topic generation. Probabilistic models, such as Latent Dirichlet Allocation (LDA)[16], model word co-occurrences within documents to infer thematic distributions. More recent techniques, such as BERTopic[17], employ clustering algorithms enhanced with contextual semantic embeddings, yielding greater flexibility and precision. Additionally, advanced neural network models, such as TopClus[18], have recently emerged, with the aim of improving the semantic coherence and interpretability of generated topics. However, the performance of these topic modeling techniques has yet to be fully validated, particularly in challenging scenarios such as social networks, where topics are highly

[1]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain. [2]Information Retrieval Lab, CITIC, Universidade de A Coruña, A Coruña, Spain. ✉email: manuel.couto.pintos@usc.es

diverse, often chaotic, and textual quality is highly variable. Furthermore, investigating topics that reflect the psychological states of social media users is of increasing research interest.

Although some comparisons among these models exist, they are often incomplete or limited to well-curated datasets with clear and well-separated topics. For example, studies often work with news corpora, which are easier to analyze. To address these limitations, we have conducted a thorough comparison using data from the eRisk collections[19–21]. This benchmark poses particular challenges due to the sparse and noisy nature of its content. The dataset contains user-generated data related to depression, offering a unique opportunity to discover latent topics expressed by individuals affected by this disorder. By massively processing thousands of social media posts and identifying key topics for psychological analysis, we can support new decision-making services (e.g., by informing new screening tools oriented towards improving the quality of life for affected individuals).

To achieve these goals, we first need to determine which model produces the most coherent topics. We consider both automatic metrics and human assessments to evaluate the quality of the extracted topics. Although automatic coherence metrics provide a rapid and easy-to-implement solution, their limitations in assessing topic quality have been well documented[14,15]. Therefore, we complement evaluations assisted by automatic metrics with rigorous validation based on human annotators. To facilitate human assessment, we have developed a software tool for topic annotation. This tool displays topic descriptions, including lists of keywords (top words) and representative texts, and incorporates a set of questions designed to help annotators rate topic coherence.

Our results demonstrate that BERTopic produces the most robust topics, excelling both in the quality of descriptive top words and in the coherence of the core documents associated with each topic. TopClus, despite being introduced in the literature as a model that outperforms BERTopic, achieved performance that was similar to BERTopic in terms of top words, but yielded poorer results in document-to-topic associations. In contrast, LDA showed the lowest performance in both aspects, highlighting the limitations of classical approaches based solely on word co-occurrence. Aside from model comparison, our experiments confirm the limited reliability of automatic coherence metrics, which show weak correlation with human judgments. This underscores the need for hybrid evaluation strategies. Furthermore, the topics extracted allow us to identify clinically relevant patterns in depression-related datasets. For example, we observe topics strongly associated with mental health struggles and self-harm, as well as socially oriented themes such as weight loss or gender identity. Such themes appear with greater intensity among depressed users. Notably, our temporal analysis reveals an increased prominence of certain critical topics during the COVID-19 pandemic, highlighting the influence of broader social events on online mental health discourse. Taken together, these findings emphasize the importance of topic modeling not only as a methodological contribution but also as a tool to support clinical practice and enrich our understanding of psychological risks in online communities.

Specifically, the contributions of this study are:

- Comparison of topic generation models: We evaluate and compare the performance of three primary approaches to topic generation (LDA, TopClus, and BERTopic) in a challenging environment characterized by unstructured and noisy data from social networks. This allows for the empirical assessment of the model best suited to such a stringent scenario.
- Validation through human evaluation: We complement automatic coherence evaluations with comprehensive human validation. To this end, we employ a platform specifically developed to assess the quality of generated topics. As a result, we have produced a dataset consisting of extracted topics and human ratings, which we make available to the scientific community. This annotated resource is publicly available at Zenodo (https://doi.org/10.5281/zenodo.15081947).
- Analysis and psychological insights from the generated topics: We examine in detail the topics derived from depression-related datasets. This allows the identification of important patterns (e.g., differences between users suffering from depression and control groups). We also leverage established clinical instruments, such as well-known symptom inventories, and Large Language Models, to deepen the interpretation of the data and extract useful knowledge. This enables the exploration of key topics, their summaries, and related data to support psychological analysis, which may be of interest to medical professionals and public institutions. For example, professionals who have not yet participated in clinical practice can exploit this technology for training purposes.

The remainder of this article is structured as follows. Section "Related work" reviews related work, including topic modeling approaches, coherence metrics, and social media analyses. Section "Topic analysis" describes the three representative models considered in this study, Latent Dirichlet Allocation (LDA), BERTopic, and TopClus, as well as the coherence evaluation strategies employed, both automatic and human. Section "Methodology" details the methodology, including a description of the eRisk dataset, the topic extraction procedure, and the human evaluation process. Section "Results" presents the results, covering the topic evaluation, a qualitative analysis of the topics generated by BERTopic, the BDI analysis, and the temporal dynamics of the extracted topics. Section "Limitations and Ethical Issues" discusses the limitations and ethical considerations, and Section "Conclusions" concludes the article.

Related work

Topic analysis models

The study of topic models has evolved significantly since its inception, with foundations rooted in classical dimensionality reduction methods. Early models, such as Latent Semantic Analysis (LSA) and Non-Negative Matrix Factorization (NNMF), employed algebraic techniques to extract latent concepts from textual collections[1,22,23]. However, both approaches were limited to the "Bag of Words" representation, disregarding the inherent sequential structure of language and the contextual relationships between words.

LDA represented a key advance in topic modeling by introducing a generative probabilistic framework capable of discovering latent topics through probability distributions[16]. Variants such as Correlated Topic Models and the Pachinko Allocation Model expanded the flexibility of topic correlation modeling[24,25], while dynamic approaches such as the Dynamic Topic Model (DTM) enabled the analysis of topic evolution over time[26].

In recent years, advances in distributed language representations have revolutionized the field of Natural Language Processing. Pre-trained Language Models (PLMs) have proven essential for improving topic model performance. Recent studies have enriched document-topic distributions[27,28] by jointly combining bag-of-words representations with contextual embeddings generated by models such as Sentence-BERT. These approaches can improve the interpretation and coherence of the topics, paving the way for the development of more advanced models.

Neural Topic Models (NTMs), based on Variational Autoencoders (VAEs), have facilitated the inference of document-topic distributions using probabilistic approximations and more sophisticated priors[29,30]. Additionally, novel strategies such as hyperbolic embeddings and optimizations based on transport distances have been explored in this area[31,32].

In parallel, non-generative, clustering-based approaches have gained popularity. Models such as Top2Vec and BERTopic combine high-quality embeddings from pre-trained models with dimensionality reduction and clustering techniques[17,33]. These methods excel in producing interpretable and coherent topics by leveraging the semantic information contained within embeddings. TopClus, on the other hand, jointly integrates document-topic and topic-word distributions using clustering algorithms[18].

Recently, alternative paradigms to VAEs have been proposed. FASTopic[34], similar to Top2Vec and BERTopic, utilizes document embeddings from pre-trained models; however, it employs optimal transport plans between document and topic embeddings to model document-topic distributions. FASTopic also models topic-word distributions via transport plans between topic and word embeddings, and ultimately optimizes both sets of embeddings by reconstructing them through these semantic relationships.

In this context, characterized by a significant development from traditional models such as LDA to contemporary neural models, we have selected three representative models from distinct families (LDA, BERTopic, and TopClus). These models support our research on topic extraction in social media and discovery of knowledge related to the psychological aspects of individuals.

Coherence metrics

The evaluation of topic quality has long been a topic of sustained interest, typically addressed through the analysis of the principal words within each topic and their overall coherence. In early approaches, human annotators were hired for this purpose; however, human evaluation proved prohibitively expensive and resource-intensive. Consequently, automated metrics were developed to assess topic quality.

In 2009, AlSumait et al.[35] introduced an automated metric aimed at identifying irrelevant or junk topics, based on statistically uniform, empty, or generic distributions. Subsequently, Newman et al.[36] proposed the UCI metric, which focuses on the co-occurrence of word pairs in a reference corpus. This metric employs Pointwise Mutual Information (PMI) as a measure of coherence.

Mimno et al.[37] later developed the UMass metric, which is based on the conditional probability of word co-occurrence within the same document. A few years afterwards, Aletras and Stevenson[38] introduced NPMI, a normalized variant of PMI, evaluated using distributional semantics and vector-space distances. Röder et al.[39] undertook a comprehensive analysis of several coherence metrics by introducing a framework that spans a broader configuration space of coherence. In that study, the authors also presented two robust new metrics, CV and CP, which combine aspects of existing measures and demonstrated improved correlations with human judgment.

More recently, Nikolenko[40] introduced the Distributed Word Representation (DWR) metrics , which evaluates the proximity of top words in semantic space, while Fang et al.[41] explored the use of embedding-based coherence measures, such as Word2Vec and GloVe. These approaches exemplify the continuing effort to exploit distributed semantic representations for improving topic coherence metrics.

Although traditional metrics have enabled the automatic evaluation of topics, they possess significant limitations. One common issue is dependency on a reference corpus; for instance, metrics such as PMI or CV require an external corpus to calculate co-occurrences, introducing potential biases if the corpus is not representative of the relevant domain. Moreover, while some metrics show moderate correlation with human judgments, they may fail to capture deeper aspects of semantics and contextual interpretability. In addition, many classical approaches overlook contextual nuances between words, making them inadequate for more complex evaluation tasks.

To address these limitations, hybrid approaches that combine automatic metrics with human review have been proposed. More recently, there has been a growing interest in using automatic evaluations based on pre-trained Large Language Models. Models such as BERT or GPT can generate simulated annotations or automatically validate topic quality, serving as a natural extension of distributional coherence metrics[42].

In our study, we employ a combination of automatic metrics and human judgments, supported by the development of a custom annotation platform. To evaluate topic coherence, human experts use a 3-point Likert scale; this choice follows[37,43], who demonstrated that simplified scales reduce annotator fatigue and improve inter-rater agreement by focusing on categorical semantic clarity. The use of both evaluation strategies ensures a robust and comprehensive approach to selecting the most appropriate topic model.

Social media analysis

Analyzing social media data has proven to be a valuable method in mental health research, owing to its potential to reveal patterns and early indicators of psychological disorders. Initiatives such as *eRisk* have provided linguistic

resources and encouraged the development of automated systems for the early detection of psychological risks. This annual event has been proposing a range of challenging tasks that have inspired teams worldwide to contribute novel and creative solutions. The participating teams work with user-generated texts published online, including those from forums, social networks, and blogs[19–21,44–46].

CLPsych is another shared-task campaign that brings together researchers from Natural Language Processing, Clinical Psychology, and related disciplines. It is oriented to develop computational tools that enhance the understanding, diagnosis, and treatment of mental disorders[47–52]. This type of evaluation-oriented events have driven the development of a wide array of methodological approaches, the progress of which has been documented in several review articles[53–58].

Among the techniques employed in this area, unsupervised methods, such as clustering, are particularly promising for analyzing social media texts and identifying patterns associated with mental health issues[59–62]. In particular, topic analysis has been exploited to investigate online communities and to distinguish subgroups among users affected by depression[6–10]. Additionally, several studies have explored the temporal evolution of topics by applying generative probabilistic models, such as LDA and its variants[11–13].

Beyond mental health, topic modeling combined with temporal and structural network analyses has been applied to study broader societal conversations on platforms like Reddit. For instance, Yang et al.[63] analyzed AI-related discourse by integrating topic evolution, posting dynamics, and user network characteristics, showing how these methods can reveal emergent patterns in online communities.

Recent contributions have underscored the relevance of language models and neural architectures for mental health applications. For example, transformer-based models have shown promise in detecting depressive content in social media[64], and recurrent neural networks have been applied for the classification of depressive messages[65]. Naseem et al.[66] have proposed an architecture to incorporate historical user data for mental health surveillance, while other authors have explored multimodal approaches. For instance, He et al.[67] introduced DPD Net, a Graph Neural Network-enhanced Transformer that integrates textual, audio, and visual modalities. Further, explainable AI methods have been leveraged to identify indicators of psychological well-being on platforms like Reddit[68]. More recently, Bao et al.[69] jointly tackled early detection and explainability using PsySym and BDISen datasets, yielding clinician-aligned symptom rationales.

Despite these advances, there is still a notable lack of research systematically investigating the application of topic analysis technologies to existing corpora for psychological risk assessment. While various studies have examined social media analysis from a broader perspective[70–74], a more focused and in-depth research is required to fully harness topic modeling for psychological monitoring tasks. The effective integration of topic analysis and social media monitoring constitutes a promising direction of research in this area.

Topic analysis

Topic models are essential tools in NLP for uncovering latent themes within large text collections. In this study, we evaluate three representative models: LDA, BERTopic, and TopClus, considering their strengths and limitations.

Latent Dirichlet allocation

LDA is a generative probabilistic model that represents documents as mixtures of latent topics, with each topic characterized by a probability distribution over words[16]. The model assumes that each document is generated by a mixture of topics, and that each word within a document is selected according to the word distribution associated with a particular topic.

From a Bayesian perspective, LDA can be interpreted as an empirical parametric Bayesian model, which represents documents as mixtures of latent topics[16]. The hyperparameters α and β correspond to corpus-level distributions: α governs the dispersion of topics within documents, while β determines the distribution of words within topics[16]. The parameter θ is a document-specific variable that defines the distribution of topics for a given document[16]. The model operates at three hierarchical levels:

- Corpus level: The parameters α and β are fixed for the entire corpus.
- Document level: For each document d, a topic distribution θ_d is generated.
- Word level: For each word w_{dn} in document d, a topic z_{dn} is chosen according to the distribution θ_d, and the word is generated based on the word distribution β of the selected topic[16].

Inference in LDA requires the computation of the posterior distribution over the hidden variables, θ and z, given the observed words, w, and the fixed hyperparameters α and β:

$$p(\theta, \mathrm{z}|\mathrm{w}, \alpha, \beta) = \frac{p(\theta, \mathrm{z}, \mathrm{w}|\alpha, \beta)}{p(\mathrm{w}|\alpha, \beta)} \tag{1}$$

However, this computation is intractable due to the complexity of integrating over all possible topic and word assignments in the marginal likelihood (the denominator)[2]. This complexity is represented by:

$$p(\theta, \mathrm{z}, \mathrm{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) \tag{2}$$

$$p(\mathrm{w}|\alpha,\beta) = \frac{\Gamma(\sum_i \alpha_i)}{\Pi_i \Gamma(\alpha_i)} \int \left(\prod_{i=1}^{k} \theta_i^{\alpha_i - 1}\right) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_n^j}\right) d\theta \tag{3}$$

As a result, numerical approximation techniques, such as variational inference or Gibbs sampling, are employed to estimate the posterior distributions[16].

Despite its effectiveness, LDA exhibits limitations in complex settings involving noisy texts. In such scenarios, the model's assumptions of term independence and simple multinomial distributions may not adequately capture intricate structures and dependencies within the data.

BERTopic

BERTopic represents a significant advancement in topic modeling by leveraging neural networks and clustering methods, moving away from traditional probabilistic generative techniques. Conventional models typically rely on bag-of-words representations, which limit their ability to capture complex semantic relationships between words[17]. In response, BERTopic introduced a modular approach organized into four distinct stages: vectorization, dimensionality reduction, clustering, and topic representation (see Figure 1).

The final stage of BERTopic is Topic Representation, where the generated clusters (c) are treated as documents. A variant of the TF-IDF weighting scheme, known as c-TF-IDF (class-based TF-IDF), is used to identify the most representative words (t) for each group. The c-TF-IDF score for term t in class c at time i is calculated as:

$$W_{t,c,i} = tf_{t,c,i} \cdot \log\left(\frac{A}{tf_t}\right) \tag{4}$$

where:

- $W_{t,c,i}$ is the final weighted score for term t in cluster c at time i.
- $tf_{t,c,i}$ is the term frequency of term t in cluster c at time i.
- A refers to the total number of documents/posts across all clusters (the corpus size).
- tf_t is the frequency of term t across all different clusters (i.e., how many times term t appears in total across the entire collection).
- the parameter i allows for modeling the dynamic evolution of topics over time.

The vectorization stage employs transformer-based models, such as SBERT, to generate advanced semantic representations that effectively capture relationships between words. These representations are richer than those produced by bag-of-words methods and enable more precise data modeling[28].

Next, the dimensionality reduction stage applies UMAP, a technique that preserves both local and global relationships in the reduced space, thereby improving clustering quality[75]. Reducing dimensionality is crucial because clustering algorithms based on distance measures degrade in high-dimensional spaces: as dimensionality increases, pairwise distances tend to concentrate and neighborhood structures become less meaningful, undermining distance-based clustering[17,76–79].

In the clustering stage, BERTopic uses HDBSCAN[80] to identify groups of similar texts, which are subsequently interpreted as topics. Finally, at the topic representation stage, the generated clusters are treated as documents, and a variant of the TF-IDF weighting scheme is used to identify the most representative words for each group.

TopClus

TopClus[18] addresses the limitations of classical models, such as the challenges in handling large vocabularies, noisy data, or high computational costs. It introduces a model that combines embeddings generated by a PLM (pre-trained language model) with a clustering algorithm to identify topics within a low-dimensional latent space. This approach comprises three main components:

- Attention-Based Document Embeddings:
  - Uses an attention mechanism to weight word embeddings generated by BERT, highlighting words most indicative of topics.
  - The embedding for each document is calculated as:

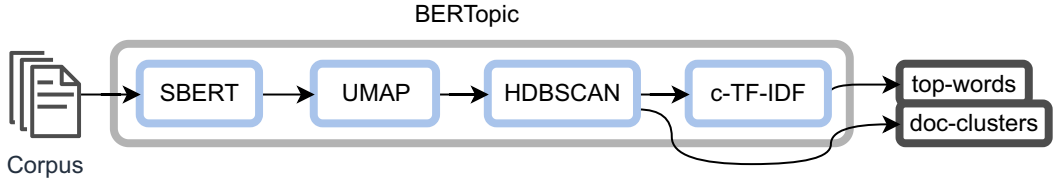


Figure 1. BERTopic workflow.

$$l_i = \tanh(W h_i^{(w)} + b) \tag{5}$$

$$\alpha_i = \frac{\exp(l_i^\top v)}{\sum_{j=1}^{n} \exp(l_j^\top v)} \tag{6}$$

$$\mathrm{h}^{(d)} = \sum_{i=1}^{n} \alpha_i h_i^{(w)} \tag{7}$$

These equations define i) l_i, an intermediate projection, ii) the coefficients α_i that represent the attention weights associated with token *i*, indicating its relative importance for topic discovery, and iii) the vector $\mathrm{h}^{(d)}$, which is the final document embedding, i.e., the semantic representation of document *d*. Here, *n* is the number of words in the document, $\mathrm{h}_i^{(w)}$ are the contextualized word embeddings generated by BERT, and $\{\mathrm{W}, \mathrm{v}, \mathrm{b}\}$ are trainable parameters of the model. Under this formulation, word- and document-level embeddings are projected into a shared latent space.

- Latent Space Generative Model:
  - Projects embeddings into a spherical latent space $Z \subset \mathbb{S}^{r'-1}$ with K clusters (topics), modeled using von Mises-Fisher (vMF) distributions. This step enhances semantics and mitigates the curse of dimensionality.
- Training Functions:
  - Distinctive Topic Clustering: Minimizes clustering loss through Expectation-Maximization (EM), promoting topic separation and balance.
  - Document Reconstruction: Ensures that topics summarize documents by minimizing the mean squared error between the original document embeddings and the embeddings reconstructed from the latent space Z weighted by the document-topic distribution.
  - Preservation of Original Embeddings: Minimizes the mean squared error of reconstructing the original contextualized word embeddings through the decoder of the model's autoencoder module.

TopClus thus combines attention, latent projection, and joint optimization to overcome the limitations of traditional models and generate coherent and meaningful topics.

Coherence metrics

Classical automatic metrics for topic coherence are computationally simple and inexpensive, but they have been widely criticized in the literature[14,15]. Therefore, to provide a more robust assessment of topics in social media, we supplemented classical metrics with additional evaluation methods involving human reviewers. Below, we describe both strategies.

Classic automatic evaluation

To assess topic quality, we employed widely recognized automatic metrics in the literature: UMass, NPMI, UCI, and CV. Due to the lack of implementations in standard libraries and because its reported performance is comparable to that of CV, we omitted CP[39]. These metrics are based on word co-occurrence and represent specific cases within the coherence framework described in[36,39]. This framework conceptualizes coherence metrics as part of a multi-stage process with four stages: i) segmentation of the top words in a topic (e.g., segmentation into word pairs and pairwise comparisons), ii) probability estimation, iii) confirmation measures (using probabilities to determine, for example, the extent to which two words are coherent in defining a topic), and iv) aggregation (e.g., averaging the coherence values of word pairs). Specifically,

- UMass[37] computes the conditional probability between words within a corpus, considering their co-occurrence across entire documents. Consequently, it is dependent on the specific corpus used for estimation.
- NPMI[38] employs a normalized approach based on fixed-size sliding windows. This yields a measure of semantic association between words that is more robust than global approaches based on word frequency within documents.
- UCI[36] also utilizes sliding windows but measures the relationship between word pairs with a simpler logarithmic metric, focusing on local co-occurrence.
- CV[39] introduces indirect confirmation measures by using cosine similarity between word vectors, which enables the capture of more abstract relationships between topic terms.

For all these metrics, the calculations are performed on specific segments of the topic's top words, and the resulting scores are subsequently averaged. This yields a single coherence value that summarizes the overall coherence of the topic.

Human evaluation

Human evaluations are considerably more expensive than automatic assessments. Nonetheless, human-generated labels constitute the gold standard or reference ground truth for estimating topic coherence. Coherence, by definition, evaluates how interpretable topics are to humans. Compared with automatic measures, human reviews offer a more reliable assessment of the quality and coherence of generated topics. To ensure robustness, we

Class	#users	percentage
Control group	1,437	84.2%
Depression group	270	15.8%
Total	1,707	100%

Table 1. Distribution of users in the dataset.

#users	1,707
avg #posts per user	12.13
Std. deviation #posts per user	33.72
minimum #posts per user	1
median #posts per user	4
maximum #posts per user	673

Table 2. Descriptive statistics of the number of posts per user.

recruited a diverse group of six annotators: two men and four women, all with university education. Five of them had a background in Computer Science, bringing technical expertise and familiarity with computational models, while the other one holds a degree in Philology, contributing valuable linguistic knowledge and sensitivity to nuances of language use. Note that the task of topic annotation does not require domain-specific expertise. The annotators only need to have rather general reading comprehension skills to assess whether the topics generated are accessible and coherent. The methodology section provides further details on the annotation procedure and guidelines followed.

Methodology

Our study aims to uncover knowledge in collections of social media posts related to mental health. To achieve this, we first determine which topic modeling approach produces the highest-quality topics. After identifying the most reliable organization of topics, we analyze the extracted topics to gain new insights into the collections.

Dataset

The dataset used in our study comprises the collections from several early risk depression tasks (editions 2017, 2018, and 2020 of eRisk)[19,20,81]. These datasets contain posts (or comments) published on Reddit by two groups of users: those who explicitly reported having been diagnosed with depression, and a control group consisting of individuals with no known diagnosis. To ensure the quality of the analysis, we removed short texts from the dataset (all posts with fewer than 17 words were removed), thereby focusing topic extraction on the most representative posts.

The dataset contains 1,707 users, of which 270 are diagnosed with depression and 1,437 belong to the control group. Table 1 summarizes the distribution of users across classes and Table 2 reports statistics on the number of posts per user. The distribution of posts per user is highly skewed: some users are very active (hundreds of posts) while other users contribute with a single post to the dataset.

Topic extraction

The topic extraction process followed these steps:

1. We fixed the hyperparameters of the three topic models to their default values. We acknowledge that LDA is known to be sensitive to hyperparameter choices and that this experimental design decision may disadvantage this model. Nevertheless, we adopted this approach to limit the number of experiments (given the high computational cost of each variant). Still, our comparison illustrates the relative performance of the models under a standard configuration that many researchers could readily reproduce using default parameter settings. This decision was made for the sake of simplicity and reproducibility, avoiding the need for additional parameter tuning that would have significantly increased the complexity of the evaluation. The number of topics was set to 50. This number was chosen based on preliminary experiments that showed that this configuration offered a good balance between topic quality and interpretability, while also keeping the annotation effort required from human evaluators at a manageable level. We aim to highlight the potential of topic analysis in this area; however, optimizing the number of topics for each method falls outside the scope of this paper.
2. We generated topics for the target collection using the three topic models.
3. We evaluated the topics using automatic metrics and human evaluation.

Human evaluation

Human evaluation was conducted by presenting the topics generated by each model to the group of six reviewers. Since we evaluated three topic analysis strategies and each produced 50 topics, each reviewer was presented with a total of 150 topics. To mitigate potential bias, the order of the 150 topics was randomized.

Each topic included a list of its top words and the five most central texts, determined by the corresponding topic analysis model. The reviewers responded to the following questions:

1. Top Words Question: "Evaluate the following list of words and globally assess its coherence in defining a topic". Response (3-point Likert scale): incoherent, moderately coherent, very coherent.
2. Topic-Name Question: "If possible, provide a name for the topic associated with the previous list of words (leave the answer blank if assigning a name is not feasible)". Response: free text.
3. Documents Question: "Evaluate the following list of texts and globally assess their thematic coherence". Response (3-point Likert scale): incoherent, moderately coherent, very coherent.
4. Document-Name Question: "If possible, provide a thematic name for this group of texts (leave the answer blank if assigning a name is not feasible)". Response: free text.
5. Global Question: "Would you say that the list of words and the group of texts are thematically coherent with each other?". Response (3-point Likert scale): incoherent, moderately coherent, very coherent.

In Figure 2, we present an example of the web form used to assess one of the generated topics. This assessment form was implemented within a web application developed specifically for storing topics and assessments in a database with a customized design. The annotation tool is open source and freely available at https://github.com/manucouto1/Topic-Quality-Assessment-Tool.

Topic analysis

After identifying the topics with the highest quality, we proceeded to analyze them to extract actionable knowledge. Since the topics generated by BERTopic were evaluated as the highest quality (see below), the subsequent analysis focused exclusively on topics produced by this model. From a knowledge exploitation perspective, incoherent topics do not provide added value; therefore, we selected only those topics with an average coherence greater than 2.5 (calculated as the mean of the coherence of the top words, the most central documents, and the coherence between the top words and the most central documents). According to the 3-point Likert scale, 1 corresponds to incoherent, 2 to moderately coherent, and 3 to very coherent.

Next, we analyzed the topics based on the presence of posts from users with depression and control users. We ranked the topics according to the proportion of texts from these two user groups and selected two sets with the ten most representative topics for the positive and control sets, respectively. Additionally, to improve the interpretability of our plots and discussion, we used a large open language model, *Llama3* (meta-llama/Llama-3.1-8B-Instruct) to assign a short descriptive name to each topic based on its top words.

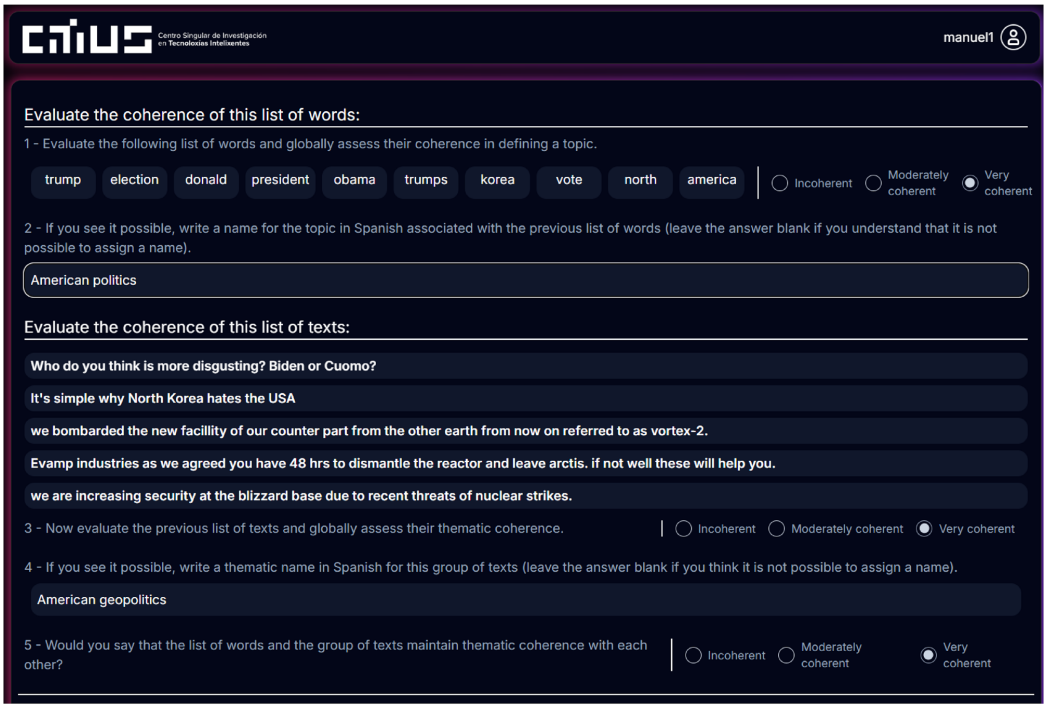


Figure 2. Web application developed for obtaining human assessments.

Automatic Coherence Metrics					Human Evaluation					
Model	**NPMI (↑)**	**UMASS (↑)**	**CV (↑)**	**UCI (↑)**	**Top Words (↑)**	κ_w	**Documents (↑)**	κ_d	**Global (↑)**	κ_g
BERTopic	**-0.030**	-5.182	**0.465**	-2.805	**2.490**	**0.376**	**2.440**	0.415	**2.350**	**0.471**
LDA	-0.056	**-2.777**	0.332	**-1.774**	1.138	0.044	1.384	**0.471**	1.150	0.133
TopClus	-0.248	-11.988	0.396	-7.198	2.429	0.327	1.945	0.364	1.606	0.395
Correlation	**0.112**	0.001	0.05	0.04						

Table 3. Evaluation of topics using automatic coherence metrics (left-hand side) and human evaluation measures (right-hand side). For each measure, the best performance is bolded. The symbols (↑) indicate that higher values of the measure correspond to better effectiveness. The average human evaluation metrics range in [1, 3], and Kappa scores (κ) measure the agreement.

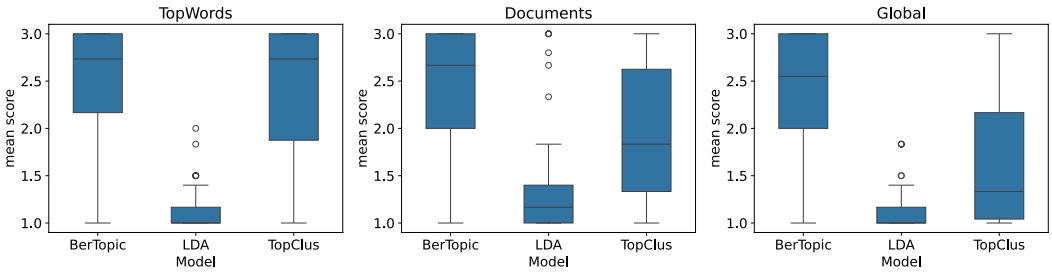


Figure 3. Boxplot of human judgments for the three topic models. The Y Axis represents the scale, from 1 to 3, used in the human evaluation.

Results

This section reports the results of the comparative evaluation of topic modeling approaches. BERTopic is considered the baseline in our experiments, as it represents a widely adopted and standardized topic modeling framework within the research community. The subsequent subsections present the human evaluation results, followed by a qualitative analysis of the topics generated by the best-performing model.

Topic evaluation

The main outcomes of the evaluation are summarized in Table 3 (automatic & human evaluation) and Figure 3 (boxplots of human assessments).

Human evaluation is often considered a more reliable assessment of topic coherence compared to automatic metrics[15,82]. To ensure statistical rigor, we computed Fleiss' Kappa (κ) to assess inter-annotator agreement among the six reviewers. This metric ensures that the observed level of agreement reflects genuine consistency rather than random variation or systematic annotator bias[83]. The computation was performed on $N = 127$ topics (out of the 150 total topics generated by the three models) that received complete ratings from all six annotators, based on the 3-point Likert scale used in the questionnaire[84,85]. The global Kappa score was approximately $\kappa = 0.50$ (ranging from 0.493 to 0.531 across the three primary coherence dimensions), indicating a moderate overall agreement among annotators. This confirms the reliability of the human assessment methodology. The values reported in Table 3 show the inter-annotator agreement (κ) disaggregated by topic model and coherence question, allowing us to assess how differences in topic quality affected the assessments. The results in Table 3 clearly indicate that BERTopic generates the highest-quality topics, achieving the highest average coherence scores (up to 2.490) and the highest level of internal agreement ($\kappa_g = 0.471$), confirming that its topics are objectively clearer and less ambiguous to interpret than the other models. Conversely, LDA exhibits extremely low agreement in Top Words ($\kappa_w = 0.044$) and Global Coherence ($\kappa_g = 0.133$), indicating that the word lists generated by LDA were statistically indistinguishable from random noise for the annotators. The moderate Document Coherence for LDA ($\kappa_d = 0.471$) reflects consensus that the document clusters were uniformly incoherent (consistent with the model's low mean score of 1.384). TopClus maintains a moderate level of agreement, but consistently below that of BERTopic (e.g., $\kappa_g = 0.395$), suggesting that its topics, while conceptually better than LDA, retained more ambiguity or noise.

In addition to coherence evaluations, reviewers were also asked to assign names to the top word lists and to the lists of central documents. Reviewers were instructed to leave these questions unanswered if they were unable to provide an appropriate name. The results were consistent with the previous findings. For example, reviewers were able to answer question 2 of the questionnaire (topic-name question) 78% of the time for BERTopic, 70.66% for TopClus, and only 8.33% for LDA. For question 4 of the questionnaire (document-name question), reviewers responded 73.6% of the time for BERTopic topics, 51.66% for TopClus, and 25.33% for LDA.

Regarding automatic coherence metrics, the results (left-hand side of Table 3) demonstrate the low reliability of these estimates. Metrics such as UMASS or UCI tend to identify LDA as the most coherent model, while only NPMI or CV show some alignment with human judgments. This finding is consistent with previous research that has reported the low reliability of automatic coherence metrics.

To provide quantitative evidence of the reliability of automatic metrics, we calculated the Spearman correlation between the average scores assigned to topics by human reviewers and by the automatic metrics. The results are shown in the last row of Table 3. Three of the metrics exhibit almost no correlation with human judgments, and only NPMI shows a slightly positive correlation.

These results demonstrate that BERTopic clearly outperforms the other models for topic organization in these collections. However, the correlations between automatic metrics and human evaluations are low, indicating substantial differences in how both approaches assess topic coherence.

To further compare the document spaces and clustering assignments produced by BERTopic and TopClus, Figure 4 presents UMAP projections of the document embedding space for BERTopic (left) and TopClus (right). To focus on confidently assigned documents, we filtered out instances with topic assignment probabilities below 0.1. This filtering reveals fundamental differences in clustering behavior: BERTopic produces 49 topics but assigns only 53% of documents with high confidence, retaining a substantial outlier cluster (topic -1) even after filtering. In contrast, TopClus generates 29 topics and successfully assigns 99% of documents, with the probability threshold eliminating nearly half of the low-quality topic candidates. The visualizations reflect distinct methodological approaches: BERTopic applies density-based clustering (HDBSCAN) to frozen BERT embeddings and then describes the resulting clusters, producing dispersed regions with many uncertain assignments. TopClus, while also using frozen BERT, learns to project both word and document embeddings into a shared latent space Z where clustering occurs, resulting in denser, more cohesive topic clusters with clearer separation and higher assignment confidence.

Qualitative analysis of generated topics

After determining that BERTopic produces the highest-quality topics, we conducted a qualitative analysis of the generated topic organizations. To this end, we created various visualizations using the procedures described in the methodology section.

Next, we examined BERTopic topics to analyze discrepancies between human reviewer evaluations and the automatic metric, paying particular attention to NPMI. Figure 5 visualizes the most coherent topics according to human judgments, filtering those with average coherence lower than 2.5. The topics are ordered from left to right by their average coherence and identified on the X-axis by the name assigned by Llama-3.1-8B-Instruct. The lower part of the plot displays a bar chart with the average coherence values assigned by the reviewers, while the upper part represents the coherence estimated by NPMI. The direction and magnitude of the bars indicate the coherence values, and the color denotes the degree of discrepancy between human judgments and the automatic metric.

Some topics highlighted in red in the upper chart, such as Time Management App, reflect a high discrepancy, as reviewers considered it highly coherent, while NPMI assigned low coherence (negative scores). Figure 6 shows the top words of this topic along with summaries generated by Llama-3.1-8B-Instruct for its central documents. According to this visualization, the top words and documents appear coherent, consistent with the human reviewers' assessments.

In the case of the topic Mental Health Struggles, the opposite occurred: reviewers assigned relatively low coherence to the top words, while NPMI considered the top words highly coherent. Figure 7 shows the top words

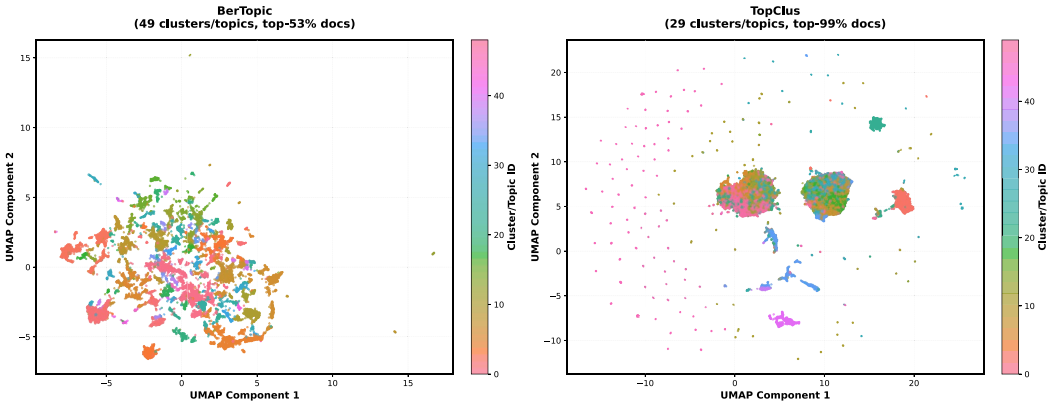


Figure 4. UMAP projections of BERTopic (left) and TopClus (right) document embeddings colored by topic assignment. Documents with assignment probability < 0.1 were filtered out, resulting in 49 topics (53% coverage) for BERTopic and 29 topics (99% coverage) for TopClus.

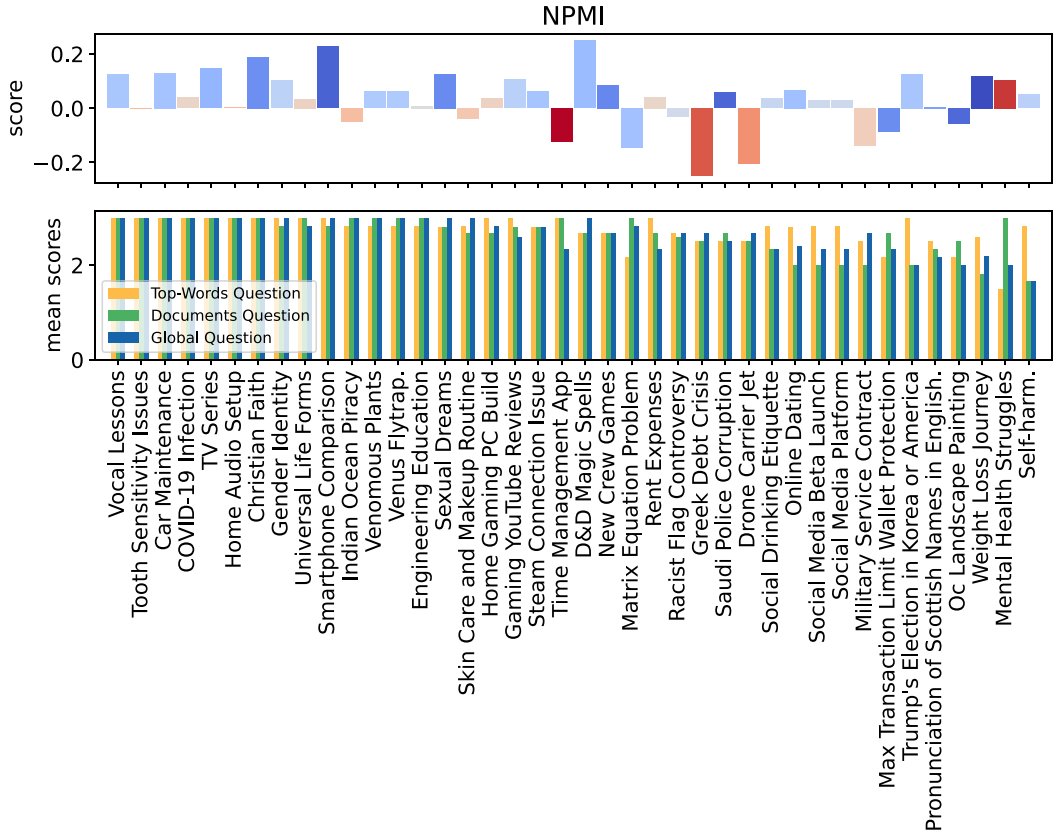


Figure 5. Average coherence levels assigned to the topics. The upper figure represents NPMI coherence estimates, and the lower figure represents the average estimates made by human reviewers.

The person has completed a bike-themed clock and is seeking feedback.

The person built their own clock and is considering taking it to show and tell at school.

A sundial showing multiple timezones simultaneously can be created using several gnomons instead of different hour markings on the ground.

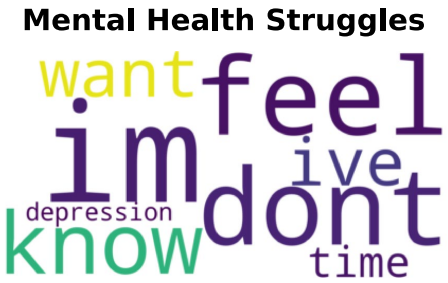
The goal is to create a sundial that displays multiple timezones simultaneously using multiple gnomons, rather than varying hour markings on the ground.

The purpose of analog clocks is compared to digital clocks, with the main point being the ease of reading on digital clocks.

Figure 6. Top words of the topic "Time Management App" and summaries produced by Llama-3.1-8B-Instruct for the topic's central documents.

and summaries of this topic. In this instance, the documents themselves appear coherent, but the top words are relatively generic (e.g., feel, want) and do not explicitly reflect the underlying theme.

Next, we analyze the posting patterns of two user groups, *Depressed* and *Control*, across the filtered topics. Figure 8 shows the total number of documents associated with each topic (first graph), the total number of users posting documents in each topic (third graph), and the distribution of documents and users from the depression and control groups across each topic (second and fourth graphs).



A 13-year-old was discharged from a mental hospital after being diagnosed with psychotic depression, triggered by severe stress and a friend's attempted suicide. They're on medication and want to make positive changes, but are struggling with a close friend who may be a trigger.

The author is seeking non-pharmaceutical ways to manage severe anxiety, having tried breathing exercises, masturbation, and exercise with little success. They are looking for alternative suggestions that have worked for others.

The author reflects on mortality, comparing life to a favorite meal - once enjoyed, the empty plate no longer bothers. They wish to feel fulfilled by life's experiences, making death easier to accept. However, they're struggling with anxiety about dying and eternal oblivion.

The writer is starting Buspar (7.5mg twice daily) to alleviate anxiety and panic issues, in addition to Citalopram (20mg daily), and is seeking feedback from others who have tried this combination.

The writer, a 17-year-old, struggles with severe depression and anxiety, experiencing insomnia, self-doubt, and poor grades, despite therapy and medication.

Figure 7. Top words of the topic "Mental Health Struggles" and summaries produced by Llama-3.1-8B-Instruct for the topic's central documents.

Some topics are clearly related to depression, such as the aforementioned Mental Health Struggles (Fig. 7) and Self-Harm (Fig. 9). The first contains a large number of documents and users, with a high proportion of content contributed by the depressed group. In contrast, the second is a more general topic concerning injuries or harm, with fewer documents and users. Additionally, the distribution of users in the self-harm topic is more balanced between the two groups.

Conversely, other topics that might not initially appear directly related to depression exhibit significant contributions from individuals experiencing this condition. This is the case for Vocal Lessons, Online Dating, Sexual Dreams, Gender Identity, and Weight Loss Journey. For some topics, such as Vocal Lessons, the situation is anecdotal, as the overall number of users posting about this topic is low (see third graph). Thus, it is not possible to draw robust conclusions about the higher proportion of users diagnosed with depression in this topic. However, the sample size for Online Dating, Sexual Dreams, Gender Identity, and Weight Loss Journey is much larger, and the high percentages of posts made by depressed individuals in these four topics clearly indicate an association between these themes and potential signs of depression.

BDI analysis

The BDI-II (Beck Depression Inventory-II) is a self-administered questionnaire designed to assess the intensity of depressive symptoms. It consists of 21 items measuring emotional, cognitive, and physical aspects; each item is scored on a scale where higher values indicate greater severity. The total score for the 21 items allows classification of depression as minimal (0–13), mild (14–19), moderate (20–28), or severe (29–63). We applied the BDI-II to the documents from the most relevant topics using prompt engineering strategies with Llama-3.1-8B-Instruct (see prompt in Figure A1, in the Appendix). The model was asked to estimate the presence of BDI-II symptoms in the documents, providing a global score that reflects the level of severity. The results are presented in Figure 10. This figure demonstrates a high proportion of documents with elevated depression estimates in the topic Mental Health Struggles. This is a natural consequence of the association between this topic and mental health issues, and it aligns with the findings of the earlier statistical analysis. Self-Harm, *Sexual Dreams*, and Weight Loss Journey also contain substantial indications of depression, while Social Drinking Etiquette includes many outliers with high scores.

Next, we closely examine summaries of documents that received the highest BDI-II scores. Figure 11 displays summaries of documents from the topic Social Drinking Etiquette. These summaries describe experiences of struggling with addiction, relapses even in supportive contexts, and narratives that encompass both internal conflict and critical or conspiratorial interpretations. This highlights the complex intersection between emotional health and social pressures surrounding alcohol consumption.

In Figure 12, we present documents with the highest BDI-II scores from the topic *Sexual Dreams*. In this case, the texts describe the intersection of sexual dreams, mental health, and traumatic experiences. Narratives of forced abstinence, reliance on nocturnal fantasies for orgasm, and conflicts arising from unresolved sexual trauma are highlighted, revealing how these phenomena reflect psychological distress. The recurrence of emotions such as guilt, frustration, and isolation underscores the connection between dreamlike sexual expression and the emotional burdens associated with social expectations, addictions, or histories of abuse. These accounts offer a critical perspective on how dreams function as mirrors of both conscious and unconscious tensions.

In Figure 13, we show documents with high BDI-II scores from the topic Weight Loss Journey, which focuses on weight-related issues. The texts reveal a complex interaction between mental health, eating disorders, and sociocultural pressures. The narratives describe experiences of chronic anorexia, with physical (alopecia, ulcers) and emotional (disconnection, loss of identity) consequences, as well as the devastating impact of negative comments about weight on self-esteem and intimate relationships. The frustration associated with conditions such as PCOS and hypothyroidism, combined with cycles of failed diets and binge eating, underscores the distress linked to the inability to meet idealized body standards. These stories, marked by guilt, hopelessness, and

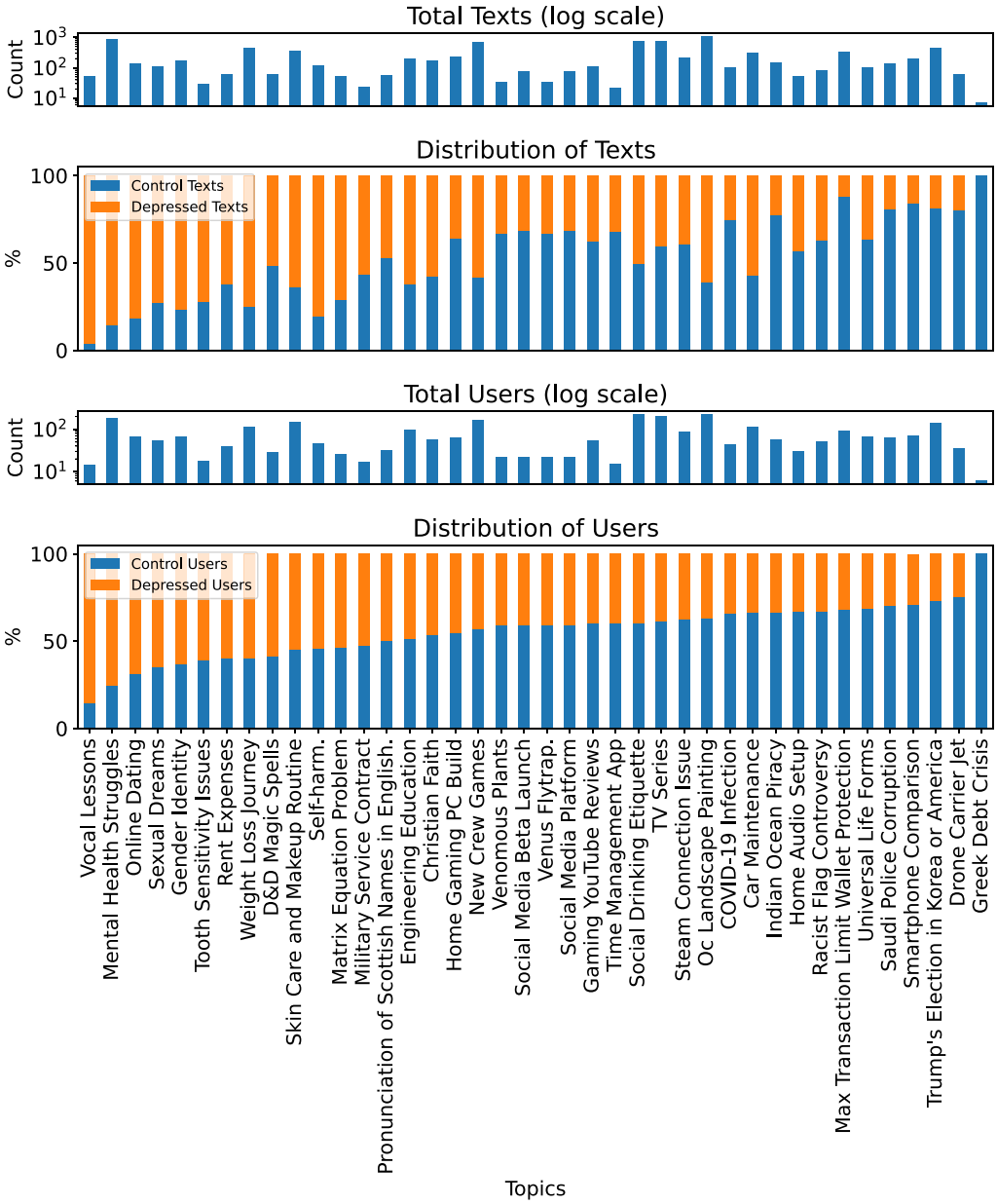
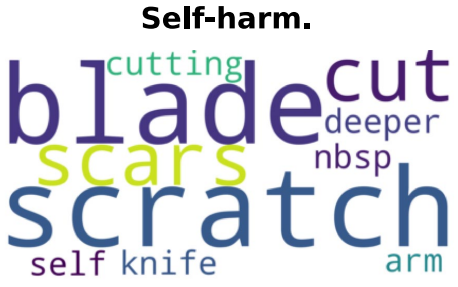


Figure 8. Total number of documents and users in each topic (first and third graphs) and distribution of documents and users in each topic (second and fourth graphs).

isolation, demonstrate that the weight loss journey not only reflects physical struggles but also serves as a mirror of profound psychological conflicts, where the body becomes a battleground for unresolved emotional tensions.

Finally, Figure 14 presents summaries of documents with the highest BDI-II scores from the topic Gender Identity. These fragments highlight trans and genderqueer experiences characterized by feelings of rejection, depression, and internal conflict, reflecting the influence of social stigmatization on the mental health of these individuals.



The author expresses discomfort with characters in the game *Injustice* who levitate instead of standing on the ground, feeling it's unnatural and wonders if one can get used to it.

A cartoon character falls 10 stories without injury, suffering only minor foot damage.

Reddit's medical community shares cases where minor physical trauma escalated to major conditions, including a minor fall resulting in a traumatic brain injury and a minor cut leading to life-threatening sepsis.

The question asks if it's possible to safely wing-suit dive from the summit of Everest, avoiding the dangers of descent.

The writer is having trouble cutting a tree with the Nattlebane, unable to damage it despite slashing at the roots.

Figure 9. Top words of the topic "Self-Harm" and summaries produced by Llama-3.1-8B-Instruct for the topic's central documents.

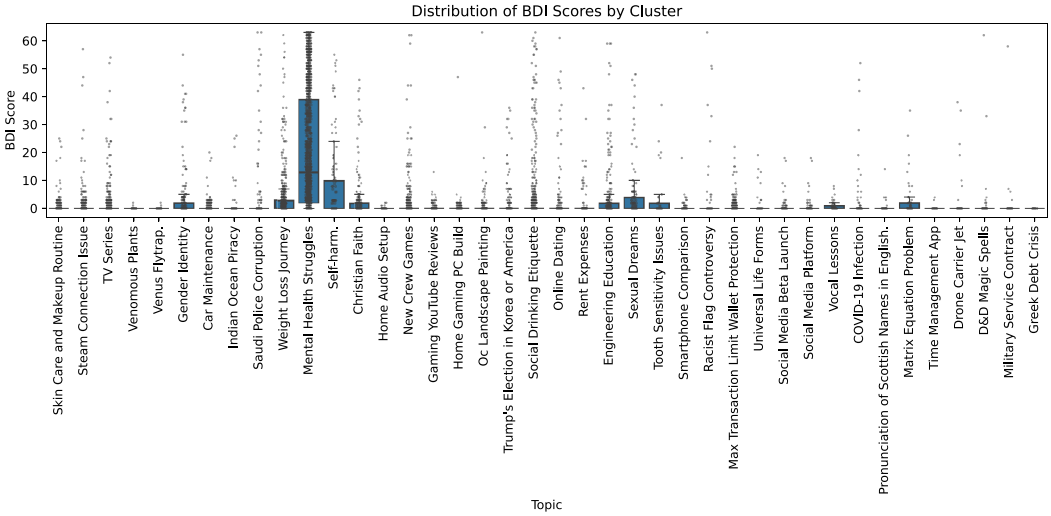


Figure 10. Results of the BDI-II evaluations on the documents of each topic.

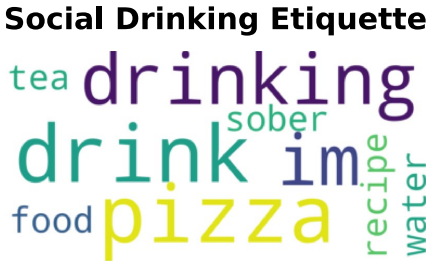
Such summaries, linked to key topics for psychological analysis, are of potential interest to medical professionals and public institutions. For example, professionals who have not yet participated in clinical practice can use this topic-plus-summarization technology for training purposes. These excerpts help elucidate the concerns and risks expressed by individuals experiencing various disorders. For preventive purposes, these new forms of screening can enable relevant public institutions to respond quickly to emerging risks in certain segments of the population. Given the high volume of posts on social media, manual monitoring is impractical, underscoring the need for automated topic organization and summarization technologies, as demonstrated here, to monitor recurring problems and their evolution.

Temporal analysis of topic dynamics

Through temporal analysis, it is possible to observe patterns in topic dynamics. Figure 15 presents, for the group of users diagnosed with depression and the five significant topics discussed earlier, the temporal evolution of the percentage of active users per topic. Notably, the percentage of active users in these topics (particularly for Mental Health Struggles) increased sharply during the pandemic years (2020 and 2021). This rise coincided with the global health crisis, and multiple studies have documented a deterioration in mental health, evidenced by an increasing frequency of posts related to depression and other psychological concerns[86].

Analysis of topics by a reasoning model

To further explore additional opportunities for exploiting the extracted topics, we conducted an exploratory (hypothesis-generating analysis) with a reasoning LLM without clinical supervision. Specifically, the model was asked i) to examine a selection of texts associated with most relevant topic for psychological analysis, and ii) to



<person> reflects on a self-destructive cycle, struggling with addiction despite having tools and resources such as a newcomer chip, numbers to call, and hobbies to engage in.

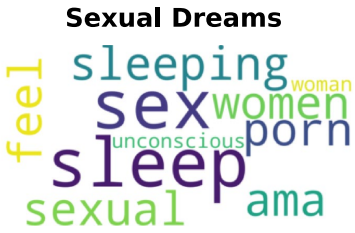
A <person> claims <enterprise> has secretly infected the population with mind-controlling nanites disguised as sugar in their Unicorn Frappuccino, part of the Northstar Initiative. The <person> has been trying to rescue those affected, but their efforts were discovered by the authorities. They now hide in a motorcourt hotel, warning others to "finish what I tried to start" and "save the world" before they are caught.

After a severe panic attack, <person> took Percocets to cope and then continued drinking, eventually relapsing due to stress from AA meetings and pressure to be truthful about their struggles. This excessive drinking was a result of guilt and shame, and <person> is now aware that their behavior will ultimately lead to self-destruction if not addressed.

<person> is experiencing a crisis due to recent accidents, reckless driving charges, and financial struggles, including overdue rent. They're considering drastic measures, but plan to call a friend for support.

A <person> struggles with heavy drinking, having lost a job at a large company, <enterprise>, due to their habit. They question whether stopping drinking will improve their life and if they can overcome past mistakes. The <person> feels anxious about living without a "distraction" and wonders if they will ever regain their former status.

Figure 11. Documents with high estimated BDI-II scores from the topic social drinking etiquette. The anonymized summaries were produced by Llama-3.1-8B-Instruct.



The writer experiences severe heartburn and describes it as extremely uncomfortable, likening it to a sensation of intense, oppressive pressure in their upper torso.

A 23-year-old male shares his unique situation: he's never masturbated, had physical contact with a woman in over a year and a half, and experienced orgasm only through nocturnal emissions.

The writer's struggle with PMO (pornography and masturbation) has been ongoing since age 12, with frequent relapses, especially after marriage. His wife's antidepressants and anxiety medications significantly reduced her sex drive, leading to his frustration and reliance on PMO. However, after sharing his struggles with his wife, they've been working to strengthen their relationship, and he's made progress in overcoming his addiction.

The writer, a female college student, expresses frustration and discomfort with her menstrual cycle, describing it as painful, messy, and unhygienic. She experiences physical symptoms such as dizziness, vomiting, and loss of appetite, as well as emotional distress from the feeling of uncleanness and inactivity. Despite improvements over time, she still struggles with the cycle's impact on her daily life and feels like she hasn't adapted to it after nearly a decade.

The writer has a history of sexual abuse and trauma, including being forced into oral sex at age 3 and 15, and experiencing ongoing abuse and manipulation in relationships. They want to give their boyfriend oral sex, but are severely triggered and anxious about it, causing them to freeze up and shut down. Despite making small progress, they are unable to overcome their fear and guilt, and worry about frustrating their boyfriend and damaging their relationship.

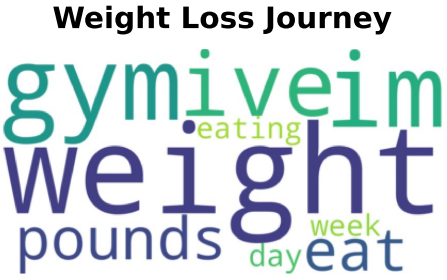
Figure 12. Documents with high estimated BDI-II scores from the topic sexual dreams.

tentatively align salient content with BDI-II symptoms. This exercise is intended to illustrate the potential of LLMs for large-scale, structured summarization, but not to provide a validated clinical assessment of a mental health condition.

For the topic Self-Harm, we selected all documents with an estimated BDI-II score of 29 or higher and submitted these texts to the LLM with appropriate instructions (see Figure A2 in the appendix). These experiments were performed with DeepSeek. The goal was to probe whether LLMs can surface plausible symptom-linked themes and produce structured summaries grounded in standardized BDI-II dimensions.

The model's synthesized summaries are reported in Figure 16, while a full trace of the model's reasoning is provided in the appendix (Figure A3). Below we outline the most salient patterns as suggested by the LLM for the Self-Harm topic (to be interpreted with caution):

- Self-Criticism and Guilt: A dominant theme in the texts is the expression of feelings of failure and disappointment. This aligns with BDI-II items related to self-blame and worthlessness, indicating a pervasive negative self-image among individuals engaging in self-harm.
- Suicidal Ideation: Explicit discussion of suicidal methods and planning indicates a high level of psychological distress. This finding is critical, as it highlights the imminent risk of self-harm escalating to suicidal behaviors.
- Loss of Interest and Avoidance: Many individuals avoid social and occupational activities due to shame associated with their scars. This avoidance reflects the broader impact of self-harm on daily functioning and quality of life.



A 26-year-old woman has struggled with anorexia for 8 years, losing friends, social life, and experiencing severe physical damage, including alopecia, stomach ulcers, and an autoimmune disorder. She feels emotionally dead, detached, and disconnected, questioning the worth of recovery and a potential new life, amidst feelings of self-disgust and loss of identity.

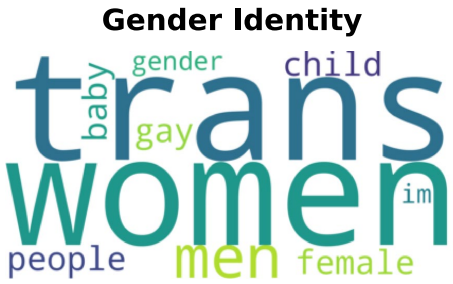
A long-term partner's hurtful comment about her weight has severely impacted her self-confidence, causing intimacy issues and a negative attitude.

A woman with PCOS, hypothyroidism, and a history of disordered eating, has tried various diets, including Whole30, but feels discouraged and unsure about her ability to lose weight and achieve her goals. She struggles with food issues, bingeing, and body image, and is currently in counseling and taking medication. Despite her efforts, she feels like giving up, but has taken the advice and encouragement from the community to heart and is making a commitment to make wiser choices and take small steps towards her goals.

A 27-year-old struggles to find identity after 10 years of battling anorexia, feeling empty, lost, and disconnected from their brain and body, with physical damage and a sense of time passing without purpose.

The writer has lost interest in things and is casually reflecting on a recent action, wearing short sleeves.

Figure 13. Documents with high estimated BDI-II scores from the topic weight loss journey.



Trans voices are dismissed on trans issues, despite being supported by science and medicine, due to cis people's lack of trust and understanding.

A trans individual, 4 months into hormone replacement therapy, is struggling with feelings of worthlessness, depression, and self-harm, seeking guidance on managing these emotions.

The author is searching for sane and relevant feminist voices in the Western world, but finds only illogical, emotionally charged, and historically inaccurate arguments, leading to frustration and ban from online feminist communities.

A 20-year-old AFAB individual identifies as genderqueer and transmasculine, struggling with inconsistent identity, depression, and anxiety, and wonders if their dysphoria could be self-inflicted due to undiagnosed Borderline Personality Disorder.

A 20-year-old AFAB individual identifies as genderqueer and transmasculine, unsure if delayed transition is harming them due to family concerns and potential self-inflicted dysphoria symptoms.

Figure 14. Documents with high estimated BDI-II scores from the topic gender identity.

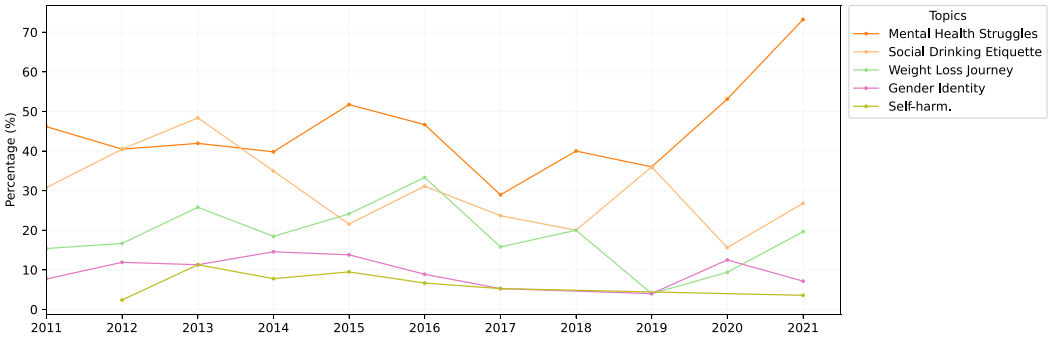


Figure 15. Temporal evolution of the percentage of active users per topic.

Introduction
The analyzed texts reflect a variety of emotional, cognitive, and behavioral patterns, with a pronounced focus on self-harming behaviors (e.g., cutting, burning) and associated difficulties. Themes of guilt, self-criticism, and suicidal ideation are consistently repeated, along with avoidance behaviors and physiological symptoms linked to self-harm. Below, the findings are contextualized within the BDI-II categories and connected to challenges in daily functioning.

Main Findings

1. Most Recurrent Symptoms

- **Self-Criticism and Guilt (Items 2, 8, 18 of the BDI-II)**:
  - *Frequency/Intensity*: Dominant in 60–70% of texts (e.g., "I feel like a total failure", "I am a complete disappointment").
  - *Context*: Frequently associated with relapse episodes (e.g., "Pathetic little cuts... broke my 10-month relapse-free streak") or fear of judgment (e.g., hiding scars from employers).
- **Suicidal Ideation (Item 9 of the BDI-II)**:
  - *Frequency/Intensity*: Explicit and repeated discussion of methods (hanging, strangulation, stabbing) and planning (e.g., "testing hanging", "how to compress the carotid artery").
  - *Context*: Current impulses are identified, though practical barriers are noted (e.g., hiding injuries from family).
- **Loss of Interest and Avoidance (Items 4, 12 of the BDI-II)**:
  - *Frequency/Intensity*: Avoidance of social activities (e.g., swimming, job interviews) due to shame over scars.
  - *Quote*: "These scars have been debilitating my self-esteem... it's hard to live a 'normal' life."
- **Agitation and Irritability (Item 17 of the BDI-II)**:
  - *Frequency/Intensity*: Frustration over the inability to self-harm (e.g., "I can't cut myself... I need this to feel real").

2. Emotional Patterns

- **Hopelessness**: Related to the perception that scars are permanent ("they will stay like this forever") and to persistent struggles with self-esteem.
- **Feelings of Worthlessness**: Reinforced by comparisons with others ("my cuts seem insignificant compared to others").
- **Shame**: Fear of exposure (e.g., during medical visits or family interactions) intensifies social withdrawal.

3. Physiological Symptoms

- **Self-Harm Injuries**: Frequent references to cuts, burns, and scars, with concerns about wound healing (e.g., "Is this burn healing well?").
- **Fatigue**: Observed indirectly through statements of stress ("too tired") and avoidance of tasks (e.g., academic assignments).

4. Cognitive Distortions

- **Catastrophizing**: Fear of losing jobs or being rejected due to scars.
- **Black-and-White Thinking**: Relapse is interpreted as total failure (e.g., "pathetic cuts... broke my streak").
- **Fixation on Pain as Validation**: Self-harm is justified as necessary for emotional survival (e.g., "it helps me feel real").

Problems and Behaviors in Daily Life

1. **Social and Occupational Dysfunction**
   - Avoidance of jobs requiring skin exposure (e.g., babysitting roles).
   - Strained relationships, such as conflicts with partners over the need to hide scars.
2. **Academic Decline**
   - Procrastination linked to emotional distress (e.g., "assignments pile up... everything is awful").
3. **Physical Health Risks**
   - Infections from untreated wounds or unsafe tools (e.g., rusty blades, chemical burns).
   - Maladaptive coping mechanisms, such as starvation as a form of self-punishment.

Conclusion
Severity Context: The symptoms align with moderate to severe depression according to BDI-II criteria, characterized by persistent self-criticism, suicidal planning, and functional impairment. Critical areas include:

- **Self-Esteem**: Widespread shame and comparisons undermine recovery.
- **Motivation**: Avoidance and fatigue interfere with daily responsibilities.
- **Safety**: Immediate risk arises from escalating self-harm and suicidal behaviors.

Impact: Daily life is heavily influenced by efforts to hide injuries, manage pain, and cope with interpersonal conflicts. Urgent attention is needed regarding coping strategies, social support, and professional intervention to address both emotional and physical risks.

Figure 16. DeepSeek-r1 analysis of documents from the self-harm topic.

- Cognitive Distortions: The LLM identified cognitive distortions, such as catastrophizing (e.g., fear of losing jobs due to scars) and black-and-white thinking (e.g., interpreting relapse as total failure). These distortions exacerbate emotional distress and reinforce self-harming behaviors.
- Physiological Symptoms and Health Risks: The analysis highlighted the physical consequences of self-harm, including untreated wounds, infections, and use of unsafe tools. These findings underscore the need for both psychological and medical care for individuals who self-harm.

These observations illustrate the complexity of depression as it is expressed in online texts and suggest themes –such as guilt, shame, suicidal ideation, and physical consequences– that may be of interest for further expert review. While these outputs should not be interpreted as clinical conclusions, they highlight the potential of LLMs to generate structured previews of large-scale text data, enabling mental health professionals and researchers to quickly identify recurring patterns and symptom-related content. In this way, LLMs may provide experts with an initial overview of evidence that likely contains traces of psychological distress, thereby guiding subsequent, clinically informed investigation.

Limitations and ethical issues

The primary objective of this study is to contribute to a deeper understanding of depressive symptoms, with the broader goal of supporting societal well-being. It is important to note, however, that this research is not intended for the clinical diagnosis of mental health conditions at the individual level.

It is also critical to acknowledge the inherent limitations and potential biases associated with the datasets and the evaluation process. As is common in social media research, the data reflect only a subset of the global population, with a strong emphasis on English-speaking users. Certain demographic groups, such as older adults who are less active on digital platforms or users with private accounts, are underrepresented. Furthermore, despite the use of Cohen's Kappa to ensure statistical agreement, human judgments are susceptible to inherent subjectivity and annotator bias. These biases can persist even in high-agreement scenarios, as labels often reflect the cultural and professional background of the annotators rather than an objective ground truth.

Another methodological constraint refers to the decision to maintain default hyperparameters for the three topic models. This is particularly relevant for LDA, which is known to be highly sensitive to parameter tuning (e.g., α and β). While parameter optimization may potentially improve LDA's performance, the sensitivity itself highlights the model's inherent lack of robustness when applied to challenging, noisy corpora, such as social media data. This contrasts with other approaches, such as BERTopic, which excel with default settings.

In our exploratory use of a reasoning LLM to analyze texts, it is important to stress that the outputs reflect model-generated hypotheses rather than clinically validated assessments. The lack of expert psychological review means that neither the scoring assignments nor the internal reasoning process of the reasoning model has been systematically validated by clinicians. The suggested associations with BDI-II dimensions are approximate, imperfect, and may contain errors, omissions, or hallucinations. Consequently, these summaries should not be interpreted as diagnostic, but rather as illustrative signals that can point to themes warranting subsequent expert review and formal validation.

Despite these constraints, our analytical framework is highly flexible and can be extended to other languages and data sources. This adaptability allows for cross-cultural investigations of how depression-related topics are expressed in various sociocultural contexts, thereby enhancing our understanding of global patterns in depressive behavior. Future research will aim to incorporate a wider range of languages and platforms, along with the development of systematic methods to identify and mitigate bias in the data.

We are acutely aware of the ethical considerations involved in analyzing social media content, particularly with respect to user privacy. To address these concerns, the study exclusively relied on pre-existing datasets that were publicly available. All methods were performed in accordance with relevant guidelines and regulations and, specifically, social media users gave their informed consent to include their data into these existing collections. All datasets used were composed solely of public interactions and were accessed in full compliance with the respective platforms' terms of service and user agreements. User identifiers (e.g., account names) were anonymized to preserve privacy and confidentiality. At no point were users contacted by us or engaged directly. This research was reviewed and approved by the Institutional Review Board (IRB) under protocol USC 65/2024.

Finally, we note that the datasets used are of limited size due to the complexities of data collection and our commitment to respecting user privacy. This study follows an observational design and does not include personal or clinical data. While such information may be critical for clinical risk assessments, it lies outside the scope of the present work. In our exploratory use of a reasoning LLM to analyze texts, it is important to stress that the outputs reflect *model-generated hypotheses* rather than clinically validated assessments. The suggested associations with BDI-II dimensions are approximate, imperfect, and may contain errors, omissions, or hallucinations. Results are also sensitive to prompt design, model choice, and dataset noise. Consequently, these summaries should not be interpreted as diagnostic, but rather as illustrative signals that can point to themes warranting subsequent expert review and formal validation.

Conclusions

This study has explored the potential of various natural language processing techniques for analyzing psychological patterns in social media texts. Identifying patterns in the language of users exhibiting signs of depression enables the development of tools to support the evaluation and monitoring of mental health, thereby complementing the efforts of professionals.

Throughout this research, we encountered challenges such as evaluating the quality of topic organizations produced by topic modeling techniques. Our experiments confirmed the limitations of existing automatic coherence metrics. By involving human judges and developing topic annotation software, we determined that BERTopic outperforms LDA and TopClus in perceived coherence. BERTopic distinguished itself by its ability to identify specific and relevant themes within noisy datasets, such as those derived from social media interactions. Our combined evaluation using both automatic metrics and human judgments confirmed the superiority of this model in terms of interpretability and usefulness for mental health research.

Beyond topic generation, this study highlighted the capability of LLMs to perform prospective analyses of user interactions on social media, particularly in relation to posts associated with psychologically relevant topics. The identification of topics linked to indicators of depression suggests the possibility of developing automated systems to guide psychologists, especially those with less clinical experience, in recognizing risk patterns. In this way, LLMs may serve as analytical assistants capable of processing large volumes of data and highlighting key information to support human intervention.

Focusing on specific findings, we identified topics strongly associated with depression, such as Mental Health Struggles and *Self-Harm*, as well as others that, although not explicitly clinical, show substantial activity among users diagnosed with depression (e.g., Weight Loss Journey or Gender Identity). The ability of text mining

models to segment and categorize these conversations opens new avenues for research and development in mental health support tools.

In future research, we intend to explore alternative topic models that may improve the coherence and accuracy of extracted topics. Additionally, we aim to investigate the application of topic extraction methodologies to other psychological disorders, such as anorexia or gambling addiction. Such efforts could further expand the utility of these unsupervised techniques in clinical practice. Interactive tools that integrate these analytical models into mental health support platforms could also be developed, thereby facilitating interventions by psychology and psychiatry professionals.

Data availability

The dataset consisting of extracted topics and human ratings is publicly available at Zenodo: https://doi.org/10.5281/zenodo.15081947. The system built for topic assessment (web application developed specifically for storing topics and assessments in a database with a customized design) is available at https://github.com/manucouto1/Topic-Quality-Assessment-Tool. The eRisk datasets used in this study are publicly available (eRisk website). Specifically, the eRisk 2017 and 2018 datasets can be obtained at https://tec.citius.usc.es/ir/code/eRisk.html (note that 2017's data was 2018's training split), while the eRisk 2019 dataset is available at https://erisk.irlab.org/2019/eRisk2019.html.

Received: 10 October 2025; Accepted: 12 January 2026
Published online: 23 January 2026

References

1. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. & Harshman, R. Indexing by latent semantic analysis. *J. Am. Soc. Inform. Sci.* **41**, 391–407 (1990).
2. Alghamdi, R. & Alfalqi, K. A survey of topic modeling in text mining. *Int. J. Adv. Comput. Sci. Appl.* **6** (2015).
3. Boyd-Graber, J. et al. Applications of topic models. *Foundations Trends Inform. Retrieval* **11**, 143–296 (2017).
4. Blei, D. M. Probabilistic topic models. *Commun. ACM* **55**, 77–84 (2012).
5. Wu, X., Dong, X., Nguyen, T. T. & Luu, A. T. Effective neural topic modeling with embedding clustering regularization. In *International Conference on Machine Learning*, 37335–37357 (PMLR, 2023).
6. Paul, M. J. & Dredze, M. Discovering health topics in social media using topic models. *PloS one* **9**, e103408 (2014).
7. Nguyen, T., Phung, D., Dao, B., Venkatesh, S. & Berk, M. Affective and content analysis of online depression communities. *IEEE Trans. Affective Comput.* **5**, 217–226 (2014).
8. Nguyen, T. et al. Using linguistic and topic analysis to classify sub-groups of online depression communities. *Multimedia Tools Appl.* **76**, 10653–10676 (2017).
9. Seo, H. & Song, M. An analysis of the discourse topics of users who exhibit symptoms of depression on social media. *J. Korean Soc. Inform. Manage.* **36**, 207–226 (2019).
10. Sik, D., Németh, R. & Katona, E. Topic modelling online depression forums: beyond narratives of self-objectification and self-blaming. *J. Mental Health* **32**, 386–395 (2023).
11. Liu, Y. et al. Monitoring covid-19 pandemic through the lens of social media using natural language processing and machine learning. *Health Inform. Sci. Syst.* **9**, 25 (2021).
12. Yu, L. et al. Detecting changes in attitudes toward depression on chinese social media: a text analysis. *J. Affective Disorders* **280**, 354–363 (2021).
13. Chandrasekaran, R., Kotaki, S. & Nagaraja, A. H. Detecting and tracking depression through temporal topic modeling of tweets: insights from a 180-day study. *Npj Mental Health Res.* **3**, 1–10 (2024).
14. Hoyle, A. et al. Is automated topic model evaluation broken? the incoherence of coherence. *Adv. Neural Inform. Processing Syst.* **34**, 2018–2033 (2021).
15. Doogan, C. & Buntine, W. Topic model or topic twaddle? re-evaluating demantic interpretability measures. In: *North American Association for Computational Linguistics 2021*, 3824–3848 (Association for Computational Linguistics (ACL), 2021).
16. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent dirichlet allocation. *J. Mach, Learning Res.* **3**, 993–1022 (2003).
17. Grootendorst, M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure (2022). ArXiv:2203.05794 [cs].
18. Meng, Y., Zhang, Y., Huang, J., Zhang, Y. & Han, J. Topic discovery via latent space clustering of pretrained language model representations. *In Proceedings of the ACM Web Conference* **2022**, 3143–3152 (2022).
19. Losada, D. E., Crestani, F. & Parapar, J. erisk 2017: Clef lab on early risk prediction on the internet: experimental foundations. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11–14, 2017, Proceedings 8*, 346–360 (Springer, 2017).
20. Losada, D. E., Crestani, F. & Parapar, J. Overview of erisk: early risk prediction on the internet. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018, Proceedings 9*, 343–361 (Springer, 2018).
21. Parapar, J., Martín-Rodilla, P., Losada, D. E. & Crestani, F. eRisk 2022: pathological gambling, depression, and eating disorder challenges. In: *European Conference on Information Retrieval*, 436–442 (Springer, 2022).
22. Paatero, P. & Tapper, U. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* **5**, 111–126 (1994).
23. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999) (Publisher: Nature Publishing Group.).
24. Blei, D. M. & Lafferty, J. D. A correlated topic model of science. *Ann. Appl. Statistics* **1**, 17–35 (2007).
25. Li, W. & McCallum, A. Pachinko allocation: Dag-structured mixture models of topic correlations. In: *Proceedings of the 23rd international conference on Machine learning*, 577–584 (2006).
26. Blei, D. M. & Lafferty, J. D. Dynamic topic models. In: *Proceedings of the 23rd international conference on Machine learning*, 113–120 (2006).
27. Bianchi, F., Terragni, S. & Hovy, D. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. Preprint at arXiv:2004.03974 (2020).
28. Reimers, N. Sentence-BERT: Sentence embeddings using siamese bert-networks. Preprint at arXiv:1908.10084 (2019).
29. Miao, Y., Yu, L. & Blunsom, P. Neural variational inference for text processing. In: *International conference on machine learning*, 1727–1736 (PMLR, 2016).
30. Srivastava, A. & Sutton, C. Autoencoding variational inference for topic models. Preprint at arXiv:1703.01488 (2017).
31. Xu, Y. et al. Hyperminer: Topic taxonomy mining with hyperbolic embedding. *Adv. Neural Inform. Processing Syst.* **35**, 31557–31570 (2022).

32. Wang, D. *et al.* Representing mixtures of word embeddings with mixtures of topic embeddings. Preprint at arXiv:2203.01570 (2022).
33. Angelov, D. Top2vec: Distributed representations of topics. Preprint at arXiv:2008.09470 (2020).
34. Wu, X., Nguyen, T., Zhang, D. C., Wang, W. Y. & Luu, A. T. Fastopic: A fast, adaptive, stable, and transferable topic modeling paradigm. Preprint atarXiv:2405.17978 (2024).
35. AlSumait, L., Barbará, D., Gentle, J. & Domeniconi, C. Topic significance ranking of LDA generative models. In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2009, Bled, Slovenia, September 7-11, 2009, Proceedings, Part I 20*, 67–82 (Springer, 2009).
36. Newman, D., Lau, J. H., Grieser, K. & Baldwin, T. Automatic evaluation of topic coherence. In: *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, 100–108 (2010).
37. Mimno, D., Wallach, H., Talley, E., Leenders, M. & McCallum, A. Optimizing semantic coherence in topic models. In: *Proceedings of the 2011 conference on empirical methods in natural language processing*, 262–272 (2011).
38. Aletras, N. & Stevenson, M. Evaluating topic coherence using distributional semantics. In Koller, A. & Erk, K. (eds.) *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*, 13–22 (Association for Computational Linguistics, Potsdam, Germany, 2013).
39. Röder, M., Both, A. & Hinneburg, A. Exploring the space of topic coherence measures. In: *Proceedings of the eighth ACM international conference on Web search and data mining*, 399–408 (2015).
40. Nikolenko, S. I. Topic quality metrics based on distributed word representations. In: *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 1029–1032 (2016).
41. Fang, A., Macdonald, C., Ounis, I. & Habel, P. Using word embedding to evaluate the coherence of topics from twitter data. In: *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 1057–1060 (2016).
42. Rahimi, H. *et al.* Contextualized topic coherence metrics. Preprint at arXiv:2305.14587 (2023).
43. Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. & Blei, D. Reading tea leaves: How humans interpret topic models. *Adv. Neural Inform. Processing Syst.* **22** (2009).
44. Losada, D. E., Crestani, F. & Parapar, J. Overview of eRisk at CLEF 2019: Early risk prediction on the internet (extended overview). *Working Notes of CLEF 2019 Conference and Labs of the Evaluation Forum* **4**, 21 (2019).
45. Parapar, J., Martín-Rodilla, P., Losada, D. E. & Crestani, F. Overview of erisk at clef 2021: Early risk prediction on the internet (extended overview).. *Working Notes of CLEF 2021 Conference and Labs of the Evaluation Forum* **1**, 864–887 (2021).
46. Parapar, J., Martín-Rodilla, P., Losada, D. E. & Crestani, F. Overview of erisk 2023: Early risk prediction on the internet. In: *International Conference of the Cross-Language Evaluation Forum for European Languages*, 294–315 (Springer, 2023).
47. Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K. & Mitchell, M. Clpsych 2015 shared task: Depression and ptsd on twitter. In: *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, 31–39 (2015).
48. Milne, D. N., Pink, G., Hachey, B. & Calvo, R. A. Clpsych 2016 shared task: Triaging content in online peer-support forums. In: *Proceedings of the third workshop on computational linguistics and clinical psychology*, 118–127 (2016).
49. Lynn, V. *et al.* CLPsych 2018 shared task: Predicting current and future psychological health from childhood essays. In: *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, 37–46 (2018).
50. Zirikly, A., Resnik, P., Uzuner, O. & Hollingshead, K. Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts. In: *Proceedings of the sixth workshop on computational linguistics and clinical psychology*, 24–33 (2019).
51. Tsakalidis, A. *et al.* Overview of the CLPsych 2022 shared task: Capturing moments of change in longitudinal user posts. In: *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, 184–198 (2022).
52. Chim, J. *et al.* Overview of the CLPsych 2024 shared task: Leveraging large language models to identify evidence of suicidality risk in online posts. In: *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, 177–190 (2024).
53. Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H. & Eichstaedt, J. C. Detecting depression and mental illness on social media: an integrative review. *Current Opinion Behav. Sci.* **18**, 43–49 (2017).
54. Chancellor, S. & De Choudhury, M. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ Digital Med.* **3**, 43 (2020).
55. Ríssola, E. A., Losada, D. E. & Crestani, F. A survey of computational methods for online mental state assessment on social media. *ACM Trans. Comput. Healthcare* **2**, 1–31 (2021).
56. Chen, X. & Genc, Y. A systematic review of artificial intelligence and mental health in the context of social media. In: *International Conference on Human-Computer Interaction*, 353–368 (Springer, 2022).
57. Crestani, F., Losada, D. E. & Parapar, J. *Early Detection of Mental Health Disorders by Social Media Monitoring: The First Five Years of the ERisk Project*, 1018 (Springer Nature, 2022).
58. Ríssola, E. A., Parapar, J., Losada, D. E. & Crestani, F. A survey of the first five years of erisk: Findings and conclusions. In: *Early Detection of Mental Health Disorders by Social Media Monitoring: The First Five Years of the eRisk Project*, 31–57 (Springer, 2022).
59. Shensa, A., Sidani, J. E., Dew, M. A., Escobar-Viera, C. G. & Primack, B. A. Social media use and depression and anxiety symptoms: A cluster analysis. *Am. J. Health Behav.* **42**, 116–128 (2018).
60. Aragón, M. E., López-Monroy, A. P. & Montes-y Gómez, M. INAOE-CIMAT at eRisk 2019: Detecting signs of anorexia using fine-grained emotions. In: *Working Notes of CLEF 2019 Conference and Labs of the Evaluation Forum* (2019).
61. De Choudhury, M., Gamon, M., Counts, S. & Horvitz, E. Predicting depression via social media. In: *Proceedings of the international AAAI conference on web and social media* **7**(1), 128–137 (2013).
62. Couto, M., Parapar, J. & Losada, D. E. Comparison of clustering algorithms for knowledge discovery in social media publications: A case study of mental health analysis. *Procesamiento del lenguaje natural* **73**, 69–81 (2024).
63. Yang, P., Han, K. & Diesner, J. Topics, temporal patterns, and network characteristics of AI-related discourse on reddit. In: *International Conference on Advances in Social Networks Analysis and Mining*, 333–344 (Springer, 2024).
64. Kerasiotis, M., Ilias, L. & Askounis, D. Depression detection in social media posts using transformer-based models and auxiliary features. *Soc. Netw. Analysis Mining* **14**, 196 (2024).
65. Kanahuati-Ceballos, M. & Valdivia, L. J. Detection of depressive comments on social media using rnn, lstm, and random forest: comparison and optimization. *Soc. Netw. Anal. Mining* **14**, 44 (2024).
66. Naseem, U. et al. Incorporating historical information by disentangling hidden representations for mental health surveillance on social media. *Soc. Netw. Anal. Mining* **14**, 9 (2023).
67. He, M., Bakker, E. M. & Lew, M. S. Dpd (depression detection) net: a deep neural network for multimodal depression detection. *Health Inform. Sci. Syst.* **12**, 53 (2024).
68. Thushari, P. D. et al. Identifying discernible indications of psychological well-being using ml: explainable ai in reddit social media interactions. *Soc. Netw. Anal. Mining* **13**, 141 (2023).
69. Bao, E., Pérez, A. & Parapar, J. Explainable depression symptom detection in social media. *Health Inform. Sci. Syst.* **12**, 47 (2024).
70. Kherwa, P. & Bansal, P. Topic Modeling: A Comprehensive Review. *EAI Endorsed Trans. Scalable Inform. Syst.* **7** (2019).
71. Curiskis, S. A., Drake, B., Osborn, T. R. & Kennedy, P. J. An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit. *Inform. Proces. Manag.* **57**, 102034 (2020).

72. Blair, S. J., Bi, Y. & Mulvenna, M. D. Aggregated topic models for increasing social media topic coherence. *Appl. Intell.* **50**, 138–156 (2020).
73. Zhao, H. *et al.* Topic modelling meets deep neural networks: a survey. Preprint at arXiv:2103.00498 (2021).
74. Laureate, C. D. P., Buntine, W. & Linger, H. A systematic review of the use of topic models for short text social media analysis. *Artif. Intell. Rev.* **56**, 14223–14255 (2023).
75. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. Preprint at arXiv:1802.03426 (2018).
76. Feldbauer, R. & Flexer, A. A comprehensive empirical comparison of hubness reduction in high-dimensional spaces. *Knowl. Inform. Syst.* **59**, 137–166 (2019).
77. Sarkar, S. & Ghosh, A. K. On perfect clustering of high dimension, low sample size data. Preprint at arXiv:1612.09121 (2016).
78. Peng, D., Gui, Z. & Wu, H. Interpreting the curse of dimensionality from distance concentration and manifold effect. Preprint at arXiv:2401.00422 (2024).
79. Pestov, V. On the geometry of similarity search: dimensionality curse and concentration of measure. Preprint at cs/9901004 (1999).
80. McInnes, L. et al. HDBSCAN: Hierarchical density based clustering. *J. Open Sourc. Softw.* **2**, 205 (2017).
81. Losada, D. E., Crestani, F. & Parapar, J. eRisk 2020: Self-harm and depression challenges. In: *European conference on information retrieval*, 557–563 (Springer, 2020).
82. Hoyle, A., Goel, P. & Resnik, P. Improving neural topic models using knowledge distillation. Preprint at arXiv:2010.02377 (2020).
83. Fleiss, J. L. Measuring nominal scale agreement among many raters. *Psychol. Bull.* **76**, 378–382 (1971).
84. Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174 (1977) (Standard reference for interpreting Kappa levels.).
85. McHugh, M. L. Interrater reliability: the kappa statistic. *Biochem. Med.* **22**, 276–282 (2012).
86. Fernández-Pichel, M., Aragón, M. E., Saborido-Patiño, J. & Losada, D. E. Personality trait analysis during the covid-19 pandemic: a comparative study on social media. *J. Intell. Inform. Syst.* **62**, 117–142 (2023).

Author contributions

MC: conceptualization, methodology, investigation, formal analysis, and writing - original draft. JP and DEL: conceptualization, methodology, investigation, formal analysis, writing-review, editing, supervision, and project administration.

Funding

The first and third authors thank the financial support supplied by the Agencia Estatal de Investigación (Spain) (PID2022-137061OB-C22; MCIN/AEI/10.13039/501100011033, Plan de Recuperación, Transformación y Resiliencia, Unión Europea-Next Generation EU), Consellería de Cultura, Educación, Formación Profesional e Universidades (Centro de investigación de Galicia accreditation 2024-2027 ED431G-2023/04 and Reference Competitive Group accreditation 2022-2025, ED431C 2022/19) and the European Union (European Regional Development Fund - ERDF). These authors also acknowledge the project "Cátedra de IA aplicada a la Medicina Personalizada de Precisión" (Cátedras ENIA, TSI-100932-2023-3); Cátedras ENIA is funded by the Ministerio de Transformación Digital y Función Pública (Secretaría de Estado de Digitalización e Inteligencia Artificial); and by the NextGeneration EU-fund. The second author thanks the financial support supplied from projects: PID2022-137061OB-C21 (MCIN/AEI/10.13039/501100011033/, Ministerio de Ciencia e Innovación, ERDF A way of making Europe, by the European Union); Consellería de Educación, Universidade e Formación Profesional, Spain (accreditations 2019–2022 ED431G/01 and GPC ED431B 2022/33) and the European Regional Development Fund, which acknowledges the CITIC Research Center.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-36339-y.

Correspondence and requests for materials should be addressed to M.C.

4.3 Publication IV: Temporal Word Embeddings for Early Detection of Psychological Disorders on Social Media

Authors: Manuel Couto, Anxo Perez, Javier Parapar, David E. Losada
Published in: Journal of Healthcare Informatics Research (JHIR) [13]
Keywords: Mental Health Detection, Social Media Analysis, Machine Learning, Word Embeddings, Temporal Word Representations, Text Mining

Journal of Healthcare Informatics Research
https://doi.org/10.1007/s41666-025-00186-9

RESEARCH ARTICLE

Temporal Word Embeddings for Early Detection of Psychological Disorders on Social Media

Manuel Couto[2] · Anxo Perez[1] · Javier Parapar[1] · David E. Losada[2]

Received: 10 July 2024 / Revised: 21 November 2024 / Accepted: 13 January 2025

Abstract

Mental health disorders represent a public health challenge, where early detection is critical to mitigating adverse outcomes for individuals and society. The study of language and behavior is a pivotal component in mental health research, and the content from social media platforms serves as a valuable tool for identifying signs of mental health risks. This paper presents a novel framework leveraging temporal word embeddings to capture linguistic changes over time. We specifically aim at at identifying emerging psychological concerns on social media. By adapting temporal word representations, our approach quantifies shifts in language use that may signal mental health risks. To that end, we implement two alternative temporal word embedding models to detect linguistic variations and exploit these variations to train early detection classifiers. Our experiments, conducted on 18 datasets from the eRisk initiative (covering signs of conditions such as depression, anorexia, and self-harm), show that simple models focusing exclusively on temporal word usage patterns achieve competitive performance compared to state-of-the-art systems. Additionally, we perform a word-level analysis to understand the evolution of key terms among positive and control users. These findings underscore the potential of time-sensitive word models in this domain, being a promising avenue for future research in mental health surveillance.

Manuel Couto and Anxo Perez contributed equally to this work

Anxo Perez
anxo.pvila@udc.es

Manuel Couto
manuel.couto.pintos@usc.es

Javier Parapar
javier.parapar@udc.es

David E. Losada
david.losada@usc.es

[1] Information Retrieval Lab, CITIC, University of A Coruña, A Coruña, Spain

[2] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), University of Santiago de Compostela, Santiago de Compostela, Spain

Published online: 22 January 2025

Keywords Mental health detection · Social media analysis · Machine learning · Word embeddings · Temporal word representations · Text mining

1 Introduction

Mental health disorders represent a critical public health challenge, affecting millions of individuals worldwide. In recent years, the prevalence of mental health conditions has shown a concerning upward direction, a trend further incremented by the COVID-19 pandemic [1]. For instance, depressive disorders impact around 280 million people globally, with severe implications including high rates of suicide and disability [2]. Early detection is thus essential, not only to enable timely intervention [3] but also to reduce the societal and economic impacts of untreated conditions [4]. In this context, the link between mental disorders and the language used by individuals has received special attention. Studies in this domain demonstrated that the manner in which people communicate their thoughts and feelings can offer valuable clues about their mental state [5]. The study of language on social media represents a crucial opportunity to identify early indicators of mental health risks [6, 7]. The proliferation of social media platforms has created vast repositories of personal expressions and interactions, offering rich data for understanding linguistic cues associated with mental health conditions [8, 9]. Moreover, social media often offers spaces for users to share thoughts and feelings, often under conditions of relative anonymity, which can encourage more open and honest communication [10].

Numerous studies have demonstrated the potential of analyzing social media content from platforms such as Twitter, Reddit, and Facebook to identify markers of mental health conditions [6]. Initiatives such as the CLEF Early Risk lab (eRisk[1]) and the Computational Linguistics and Clinical Psychology workshop (CLPsych[2]) have supported this type of research, providing resources such as task definitions, datasets, and evaluation methodologies for various mental disorders [11, 12]. Seminal studies aiming to detect mental health issues through social media analysis relied on examining the entire history of a user's thread of publications. Alternatively, due to the dynamic nature of social media and the rapid evolution of mental health conditions, a promising avenue, known as early risk detection (ERD) [13], aims to facilitate the timely identification of individuals at risk.

To address this need, the eRisk campaign introduced a pioneer ERD task in 2017, focused on signs of depressive disorders [13]. ERD models differ from conventional methods by requiring a sequential analysis of user-generated posts, thus simulating how content becomes available in real scenarios. ERD models assess risk iteratively, updating their predictions with each new data chunk and emitting alerts when a significant risk is identified. Unlike conventional detection frameworks, the ERD task emphasizes the importance of temporal aspects in social media activity. Due to the popularity of this framework, the eRisk's organizers have since extended this idea to other disorders, including anorexia, gambling addiction, and self-harm [14–17]. These

[1] https://erisk.irlab.org/

[2] https://clpsych.org/

tasks require methods that can incrementally identify early risks. For this reason, eRisk datasets consist of training and test splits that present chronologically ordered posts segmented into temporal chunks.

This article introduces a novel approach to detecting temporal linguistic changes in social media users and exploiting them as potential indicators of mental health risks. Our framework permits the analysis of language variations in individuals, and we track these variations using different datasets related to four mental health conditions. The main hypothesis of our work is that the progression of a mental health condition is often accompanied by alterations in linguistic patterns. Through our analysis, we aim to reveal patterns in language evolution that may signal changes in mental health status. For instance, an individual without psychological issues might commonly associate the word "life" with positive or neutral expressions, such as "good" or "going on." However, if the same individual starts to experience indicators of mental health problems, associations with "life" may shift towards more negative contexts, appearing with terms like "ruined" or "miserable." These shifts from positive to negative contexts, which may occur gradually or suddenly, can offer insights into the temporal dynamics of emerging mental health issues.

To facilitate our exploration, we introduce the concept of temporal word embeddings and adapt it for the early detection of mental health issues. While temporal embedding models have traditionally been used to capture long-term linguistic and cultural shifts [18–20], our proposal is unique in that it applies these models to detect linguistic changes indicative of mental health risks. Specifically, our methods estimate the magnitude of change in word usage over defined time intervals. These time-aware word movements are employed as features for training classification models. The movement of a word is calculated as the distance between its temporal embedding at a given time point and a static reference embedding, representing the general or baseline usage of that word. In this study, we investigate two temporal embedding methods: (i) *Temporal Word Embeddings with a Compass* (TWEC) [21], which builds on the word2vec architecture to generate unique embeddings per word, and (ii) *Dynamic Contextualized Word Embeddings* (DCWE) [22], which leverages pre-trained models like BERT [23] to create contextualized word embeddings. Our experiments utilize the eRisk benchmark datasets from 2017 to 2022, covering a diverse range of mental health conditions, and our results indicate that temporal word representations hold significant potential for tracking the progression of these conditions. We evaluate these methods using the eRisk benchmark collections from 2017 to 2022. The experimental results suggest that temporal representations have the potential to detect the evolution of mental health conditions. Our study offers, thus, a promising avenue for understanding how changes in language usage over time can reveal early indicators of mental health concerns.

This paper extends our previous research, which explored the relationship between language usage and its associations with depression on social media [24]. Building on these foundations, we further explore this idea to cover a wider scope, examining 18 distinct datasets across four mental disorders. This expanded set of experiments allows us to study deeper about how word evolution can improve our understanding of the linguistic markers associated with mental health. In summary, our study advances the field in the following ways:

- We build a framework for studying language evolution on social media, evaluating the performance of our temporal-aware methods on 18 different datasets related to four mental illnesses.
- Our approach can naturally incorporate base classifiers of different natures, and we perform a comparative analysis to understand which type of classifier best handles word movements over time.
- Through a detailed case study, we extract those words that most influence the classification decisions for each disease and illustrate the role those keywords may play in differentiating positive and control users.

The remainder of the paper is structured as follows. Section 2 provides a review of related work about computational methods for mental health analysis, social media monitoring, and temporal modeling. Section 3 presents our proposed time-aware framework and the two variants implemented. In Section 4, we describe the collections used, the target task, the experimental settings, and the results of our temporal methods. Section 6 presents a keywords evolution analysis. Section 7 explores alternative distance metrics and their impact on early risk detection. Finally, Section 8 discusses the main conclusions derived from this study and proposes future lines of work.

2 Related Work

2.1 Computational Methods for Mental Health Analysis

Researchers in Medicine and Linguistics have extensively studied the relationship between mental health and language use. In particular, several works focused on language as an indicator for detecting mental health issues [6, 7]. Social media platforms, such as Twitter, Facebook, and Reddit, have become valuable data sources in this area [25–27]. Early work by De Choudhury et al. [8, 27] pioneered automatic depression detection on social media, extracting features from the area of Psychology, such as high self-attentional focus [26], negative emotions [27], and the use of absolute terms like "always" or "never." Building on these insights, other researchers have explored various features, including linguistic style, emotional expressions, semantic or lexicon-based variables, social network properties, and posting activity [28, 29]. The popularity of this research line has motivated the creation of shared tasks that promote the exploration of mental health topics and online behavior. Two of the most popular campaigns in this field are (i) eRisk, under the CLEF evaluation forum, and (ii) CLPsych, organized by the Association for Computational Linguistics (ACL), both of which are oriented towards linguistic markers and computational strategies in mental health analysis. Our work is closely aligned with the eRisk tasks. We, therefore, will focus our discussion on this evaluation benchmark.

The eRisk organizers have designed different tasks focusing on specific social media risk assessment aspects. A strong point of the eRisk framework was the introduction of challenges beyond binary classification, simulating real scenarios and the complexities associated with mental health-related content in social media. Specifically, our research is related to the eRisk's early risk detection task [13]. This task pioneered

the introduction of a dataset designed to process social media content in a sequential manner [30]. In this task, ERD models are required to analyze users' posts sequentially, update the estimated risk levels over time, and emit alerts when any mental health risk is detected. To facilitate this sequential approach, eRisk datasets segment the user's list of posts into ten chunks, with each chunk representing 10% of the user's posting history. Since 2017, this task has expanded to four mental health conditions: depression, anorexia, gambling addiction, and self-harm, with over 500 different ERD methods submitted by more than 50 research groups [11, 13–17, 31].

In the initial eRisk editions, researchers employed traditional machine learning techniques, such as bag-of-words [32], topic modeling [32], standard classifiers [33], and neural networks [34]. These approaches were often combined with feature extraction methods like depression lexicons, sentiment analysis, and emotional or semantic analysis [26, 35–37]. Building on these works, in 2019, [26] introduced the SS3 classifier, a model incorporating three key features: incremental classification, support for early classification, and explainability. SS3 maps words to depressive categories by estimating the confidence that a word belongs to these categories, demonstrating strong performance on the eRisk 2017 early depression dataset. Recent eRisk editions have shifted towards transformer-based models, with participants adopting architectures like BERT [38] to improve classification performance [39, 40]. In all these works, a key feature was the strategy implemented to determine the timing for emitting the final prediction. Most approaches opted to identify a user at risk once the model's output exceeded a predefined threshold. Alternatively, some teams incorporated a set of criteria that considered the volume of posts analyzed [41] or changes in risk level observed over recent days [26].

2.2 Temporal Word Embeddings (TWE)

Neural language models are widely adopted for capturing semantic features from textual data, typically generating vectorial representations of words, known as word embeddings. Word embeddings position words appearing in similar contexts closely within a multi-dimensional space. Traditional word embedding architectures, like Word2Vec [42] and GloVe [43], are static, offering a single, global embedding per word. This approach has two primary limitations [22], as static architectures do not capture how the meaning of a word varies across (i) linguistic contexts (i.e., to model polysemy) and (ii) extralinguistic contexts (i.e., to model temporal variations).

Contextualized embeddings derived from models such as BERT [38] and T5 [44] addressed the first limitation, achieving state-of-the-art results across NLP tasks. Moreover, a growing body of research focused on the second limitation, creating approaches known as Temporal Word Embedding Models (TWE), also known as diachronic models. However, modeling embeddings over time presents unique challenges. Given the stochastic nature of neural networks, models trained at different time intervals produce vector spaces with incompatible coordinate systems. Consequently, aligning temporal models becomes a prerequisite before performing a comparative analysis, which introduces both computational and performance costs.

Static TWE approaches have employed various *alignment* strategies. A common method divides data into temporal chunks, generating independent embeddings for each chunk and aligning them in a later step [18–20, 45]. For instance, [18] calculated the artificial rotation between adjacent vector spaces to identify semantic shifts, while [45] used strategic neural network initializations. These alignment methods, however, often struggle with limited data per time slice. In contrast, other approaches trained temporal embeddings within a unified space, leveraging all available data, and improved results on smaller datasets [19, 20]. For instance, [19] utilized regularization techniques to minimize changes in embeddings over time, while [20] applied probabilistic methods to maintain vector continuity. However, these studies tend to suffer from scalability and efficiency issues. The Temporal Word Embeddings with a Compass (TWEC) model [21] represented a notable advancement in this area by eliminating the need for explicit alignment across different time slices. TWEC uses atemporal vectors as a reference (*compass*) when training the temporal representations, ensuring that all temporal representations are aligned within the same coordinate system.

More recent work explored language dynamics with contextualized embeddings [46, 47]. Pre-trained transformer models facilitate the training of contextualized temporal embeddings, though they often underperform compared to static embeddings in language evolution tasks [48]. Addressing this issue, Dynamic Contextualized Word Embeddings (DCWE) [22] leveraged pre-trained language models across time intervals, integrating temporal and social information. This approach demonstrated that incorporating temporal information not only improves the ability to track linguistic evolution but also improves performance on downstream tasks, such as sentiment analysis. Our work extends the applications of temporal-aware models to study the linguistic variations in social media users suffering from signs of different mental health conditions.

3 Method

The main idea behind our models is that the progression of mental health issues often manifests through changes in language use. As individuals experience and express new emotions, thoughts, and feelings, these internal changes are reflected in their written communications. Consequently, individuals suffering from mental illnesses may vary their use of words, with certain terms being employed in new contexts that were previously uncommon. Below, we introduce the structure of our framework to capture these linguistic changes and explain the different variants implemented.

Our framework calculates two types of word embeddings: *reference* and *dynamic*. Reference embeddings serve as a static baseline, capturing the general, neutral usage of words. In contrast, dynamic embeddings are specific to each user and time interval. For each user and time period, we generate dynamic vector spaces, allowing us to measure changes in word usage across different temporal chunks. These changes are quantified by calculating the distance between dynamic embeddings and their reference vectors. The resulting patterns of word movements across intervals are then used as features for model training and during the classification process. This methodology is based on the hypothesis that the language patterns of individuals experiencing mental health

issues will exhibit more pronounced and irregular movements. Our framework consists of three main phases: (i) construction of reference embeddings, (ii) construction of dynamic embeddings, and (iii) calculation of the magnitude of word movements. The first two phases depend on the selected temporal model. The last one (computation of word movements) is model-independent and remains the same regardless of the temporal variant.

Initially, we construct a global reference space, denoted as R. This global space is atemporal: it considers the entire training corpus across all time chunks. To compute R, we train a model to generate a reference vector for each word w in the vocabulary V of the training corpus, with each word vector in this space represented as $R(w)$. After constructing the reference space, we split the user-generated posts into n temporal chunks. Each chunk is defined as C^{t_j}, representing the temporal space for that interval, where $1 \leq j \leq n$. Within our framework, the vector $C^{t_j}(w)$ represents the specific usage of word w at the time t_j. The methodology for developing both the reference and temporal space is dependent on the specific embedding model. In this study, we explore two distinct approaches utilizing both static and contextualized embedding architectures.

Next, our framework measures the magnitude of temporal word shifts within the user-generated posts of each temporal chunk, C^{t_j}. We do this by calculating the cosine distance between each word in chunk t_j and its representation in the global *reference vector space*, $R(w)$. We denote this distance as $\delta_w^{t_j}$, which quantifies the *word shift* or movement of word w at time t_j from its reference usage. For simplicity, we will refer to these *word shifts* as *deltas*. Larger delta values indicate greater deviation from the reference, suggesting notable changes in the meaning or usage of a word over time. If a word is absent in a specific chunk C^{t_j}, its delta is set to zero, indicating no observed change in usage. Formally, the movements are defined as follows:

$$\delta_w^{t_j} = \begin{cases} 1 - \frac{C^{t_j}(w) \cdot R(w)}{\|C^{t_j}(w)\| \|R(w)\|} & \text{if } w \text{ occurs in chunk } t_j, \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Alternative vector distance metrics can also be integrated into our temporal models. An analysis of these variants is provided in Section 7.

These deltas can be represented in a matrix that captures changes in an individual's language use. For each user u_i, a unique matrix Δ_{u_i} is constructed. The delta matrix is constructed as follows:

$$\Delta_{u_i} = \begin{array}{c} \\ t_1 \\ \vdots \\ t_j \\ \vdots \\ t_n \end{array} \begin{array}{c} \begin{array}{ccccc} w_1 & \cdots & w_k & \cdots & w_m \end{array} \\ \begin{pmatrix} \delta_{w_1}^{t_1} & \cdots & \delta_{w_k}^{t_1} & \cdots & \delta_{w_m}^{t_1} \\ \vdots & & \vdots & & \vdots \\ \delta_{w_1}^{t_j} & \cdots & \delta_{w_k}^{t_j} & \cdots & \delta_{w_m}^{t_j} \\ \vdots & & \vdots & & \vdots \\ \delta_{w_1}^{t_n} & \cdots & \delta_{w_k}^{t_n} & \cdots & \delta_{w_m}^{t_n} \end{pmatrix} \end{array} \tag{2}$$

In this user matrix, rows correspond to the n temporal chunks while columns represent the movements for each word w_k in the vocabulary (i.e., from the total m words in the vocabulary). Each row of the user matrix stores the word shifts for a distinct temporal chunk, t_j. For instance, the row for time interval t_j (represented as $\Delta_{u_i}[t_j]$) captures all the word shifts that occurred for the user u_i during that time interval t_j.

To study how tracking word evolution offers insights within the mental health domain, we exploit these temporal representations to address a binary classification task oriented to distinguishing between positive (i.e., users with mental health issues) and control individuals. For this purpose, we explore different machine learning classifiers, using the deltas from each chunk as features. Specifically, we treat each row of a user's delta matrix as an independent data sample. During the training phase, each row is labeled as 0 (control) or 1 (positive) based on the mental health status of the corresponding user. In the inference phase, the classifier processes user chunks sequentially (in chronological order). If the classifier's decision for a chunk is negative, the next chunk is processed. If the decision is positive, we emit an alert for that user. Finally, if the classifier reaches the last temporal chunk and the decision is still negative, the user is ultimately classified as negative. Further details regarding the classification process are detailed in Section 4.2.

3.1 Construction of Reference and Temporal Spaces

This subsection describes our approach to constructing word representations by adapting temporal models following two variants.

3.1.1 Temporal Word Embeddings with a Compass (TWEC)

Our first variant adapts a Word2vec-based temporal model, TWEC [21]. TWEC uses atemporal vectors as reference points (i.e., a compass) to guide the training of time-specific word representations. This model aligns well with our framework, allowing us to employ the compass as a reference vector space. TWEC [21] is based on the assumption, also noted by [49], that most words retain stable meanings over time. For this reason, the words with notable semantic shifts tend to co-occur with more stable words. In our study, we apply this idea to the mental health domain, hypothesizing that words reflecting personal experiences (e.g., *life, feel*) will co-occur with words whose meanings remain consistent (e.g., *family, home*). To that end, using the Word2vec architecture, TWEC employs a dual representation of words: input and output embeddings. TWEC keeps the output embeddings frozen across all temporal intervals and updates input embeddings for each time period. The frozen layer acts as an atemporal compass, ensuring all temporal models share the same coordinate system. In the original study [21], TWEC's compass was trained on an entire dataset, followed by segmentation into temporal periods for the creation of specific temporal spaces. In our case, however, the eRisk datasets present a unique challenge: user posts are provided sequentially rather than as a complete history, so we do not have access to the full publication history during inference.

To address this issue, our adaptation of TWEC processes chunks sequentially, allowing for early alerts based only on the available chunk data. The method, thus, does not rely on information from future chunks. Figure 1 illustrates our TWEC adaptation process. First, we train the Word2vec CBOW model on the training dataset to create the output weight matrices (R). These act as our reference space, capturing the general use of language across all users. After R is calculated, we generate temporal models for each user and time period ($C_{ui}^{t_j}$), where u_i is the user and t_j is the time chunk. To that end, we construct each temporal model by training on user-specific chunk data while keeping R embeddings constant to maintain a unified coordinate system. This adapted approach is replicated during inference for test users, ensuring temporal models also remain aligned.

3.1.2 Dynamic Contextualised Word Embeddings (DCWE)

An alternative approach for constructing reference and temporal models exploits contextualized word embeddings by adapting the DCWE method [22]. This approach leverages transformer architecture to produce embeddings sensitive to context. DCWE includes two main components: a dynamic module and a contextualizer, as depicted

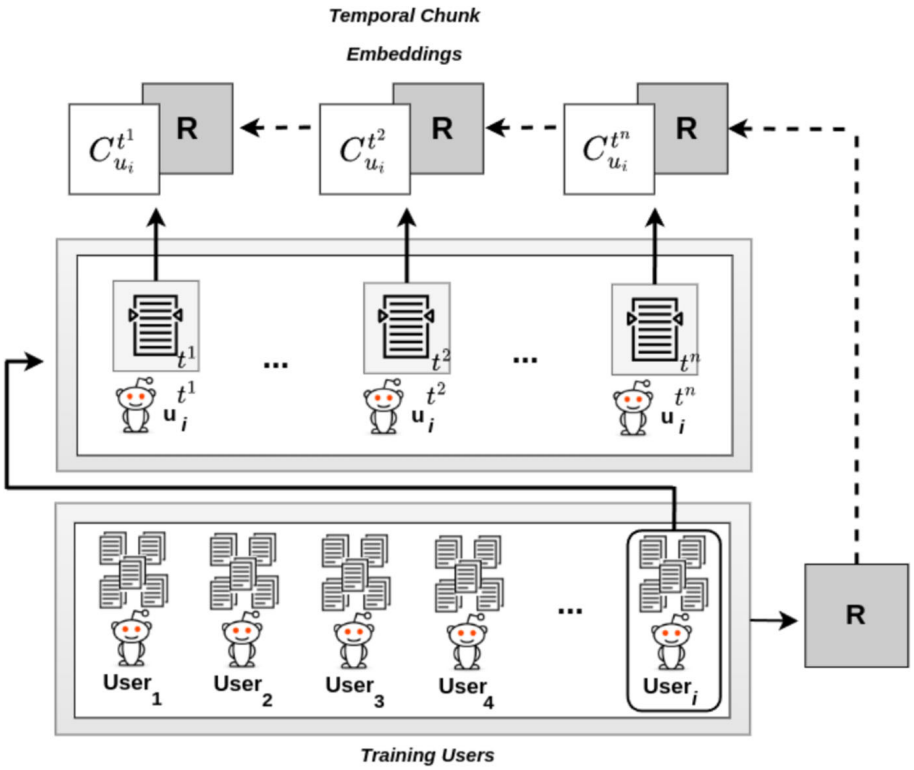
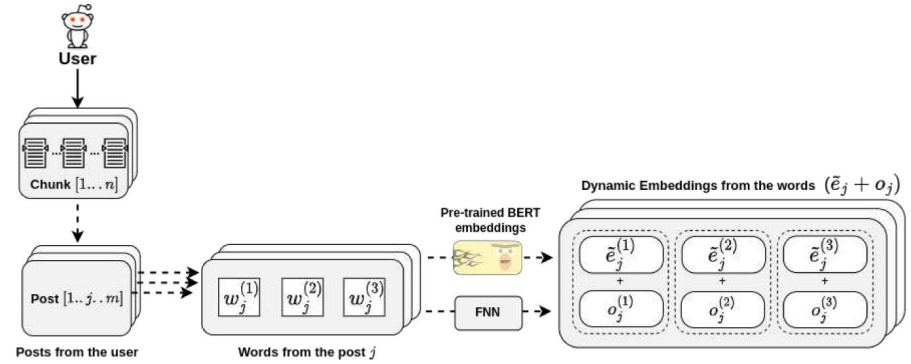


Fig. 1 TWEC adaptation for our method. The temporal chunk embeddings $C_{u_i}^{t_j}$ are trained over each time period. Temporal embeddings use the atemporal compass as a reference, R

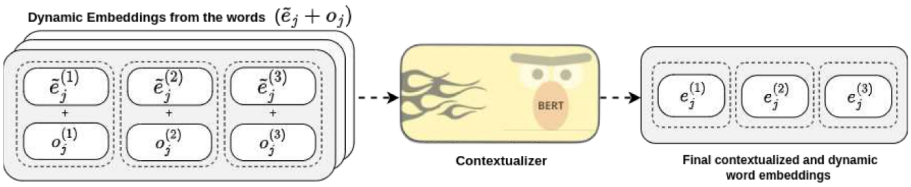
in Fig. 2 (blocks a and b, respectively). In this method, words are first mapped to dynamic representations that are then contextualized. For both components, we use the BERT$_{BASE}$ with 12 attention layers and a hidden layer of 768 dimensions. The embeddings generated through this process are denoted as e_j^w, corresponding to word w at time interval j.

The dynamic component introduces a vector offset, o_j^w, which varies based on the word and its temporal context. The offset is produced by a time-specific feed-forward neural network (FNN). To ensure gradual transitions between embeddings across time periods, the model applies a random walk prior on o_j^w, promoting gradual shifts in dynamic embeddings over time. As a result, embeddings for words across far time intervals exhibit larger differences. These dynamic offsets are then combined with static embeddings, $\tilde{e}_j^w$, initialized with BERT's pre-trained vectors, integrating temporal evolutions into the contextualized embeddings. The concatenation of the static embeddings, $\tilde{e}_j^w$, with the temporal offsets, o_j^w, forms the dynamic component of our embeddings ($\tilde{e}_j^w + o_j^w$). Finally, these dynamic embeddings are contextualized through BERT (component b in the figure), using Masked Language Modeling (MLM) as the training objective. This produces the final *dynamic and contextualized* embeddings. We note that, due to the maximum length limit of BERT, we need to split the chunks into individual posts, leading to different dynamic word embeddings for each word appearing in each individual publication.

In our scenario, to compute temporal word shifts, we need to compare the DCWE temporal embeddings against a reference vector space R. To do so, for constructing R, we utilize BERT's pre-trained static embeddings, assuming these embeddings derived



(a) Dynamic component



(b) Contextualizer component

Fig. 2 DCWE implementation of our variant. For more details of the components, please see [22]

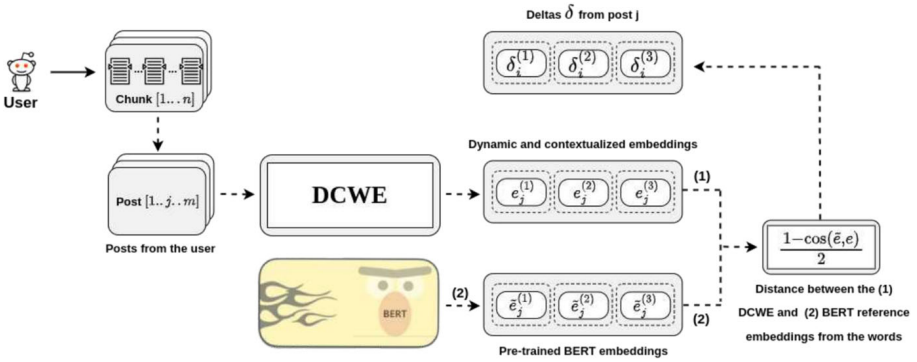


Fig. 3 Calculation of the word shifts of each time period after the DCWE model is trained. As we have a different embedding for each word occurrence, we average all these embeddings to quantify each word's shift

from extensive document collections serve as a stable reference. The DCWE temporal embeddings are compared to this reference space at the post level, as shown in Fig. 3. Thus, the word shifts (deltas) are calculated for each distinct post.[3] Finally, the shifts for each chunk are then obtained by averaging the distances for words appearing within that interval of publications.

3.1.3 TWEC vs DCWE: Main Differences

A primary distinction between DCWE and TWEC lies in their base architectures. DCWE leverages transformer-based architectures, which generate dynamic word vectors that adapt based on the context. TWEC employs Word2Vec, which produces static word representations. In terms of training methodology, TWEC first trains a "compass" on the entire document corpus, refining each temporal model to capture time-specific properties. In contrast, DCWE dynamically constructs embeddings during inference, capturing temporal features for each interval. Another notable difference in our work is the source corpus for constructing the reference space, R: TWEC's R is computed from eRisk training data, while DCWE uses BERT's pre-trained embeddings, derived from a more diverse and extensive corpus.

In terms of computational costs, we used Gensim for TWEC and PyTorch for DCWE, with PyTorch supporting GPU acceleration, giving DCWE a computational advantage. Additionally, DCWE incorporates vocabulary filtering as part of its default training process, while TWEC processes the full vocabulary. Consequently, TWEC generates a larger feature set, with a dimensionality of around 50,000 words as features, compared to approximately 9000 words for DCWE. This discrepancy not only increases TWEC's computational requirements but also introduces higher complexity and dimensionality, which can complicate the integration with other models. The characteristics of each model are summarized in Table 1.

[3] If a word does not occur in any of the posts of a chunk, we assume no movement. Thus, we set its shift value to zero (0).

Table 1 Comparison of DCWE and TWEC temporal models

Method	Base model	GPU acceleration	Vocabulary size
DCWE	BERT	✓	Medium
TWEC	Word2Vec	✗	High

4 Experiments

This section explains the experimental framework and results obtained in the evaluation of our proposed approaches. First, we describe the collections used to train and test the classifiers. We also present the evaluation metrics, which assess for both accuracy and delay in classification decisions. Further details on implementation are provided in Section 4.2.

4.1 Datasets and Evaluation Metrics

4.1.1 Datasets

The datasets utilized in our study were created in the eRisk editions from 2017 to 2023 [11], covering a total of 18 distinct datasets. These collections were designed to research the early detection of mental conditions, including four primary disorders: depression, anorexia, gambling addictions, and self-harm. Each dataset was structured to support sequential analysis of Reddit posts from users categorized as either *positive* (suffering from a mental health condition) or *control*. The users' posts are organized chronologically and segmented into ten chunks, with each chunk representing 10% of an individual's posting history. The dataset statistics are summarized in Table 2. In total, the collections have 1089 positive cases and 12,115 control users. This class imbalance between positive and control users reflects real-world scenarios, which typically have more control users compared to positive users. Notably, the 2019 self-harm and 2021 gambling datasets lack training data and, as such, were excluded from our experiments. In the future, we are planning to do cross-domain experiments where, for example, we can deploy depression-trained classifiers to detect early signs of self-harm. In any case, this type of transfer learning experiment was left out of the scope of the current study.

4.1.2 Evaluation Metrics

To evaluate performance, we adopt the official eRisk evaluation metrics, which include traditional classification metrics and time-sensitive measures. More specifically, we report the standard F_1 metric, a traditional measure that balances between precision and recall, and *Early Risk Detection Error* (ERDE), a metric introduced by Losada and Crestani [30, 31], which evaluates both the accuracy of predictions and the delay taken to emit the alert. Given d, the binary decision taken by a system, and k, the delay

Table 2 eRisk datasets statistics (depression, anorexia, self-harm, and gambling)

Depression						
Edition	2017		2018		2022	
	Train	Test	Train	Test	Train	Test
Depression	83	52	135	79	214	98
Control	403	349	752	741	1493	1302
Total	486	401	887	820	1707	1400
Anorexia						
Edition	2018		2019			
	Train	Test	Train	Test		
Anorexia	20	41	61	73		
Control	132	279	411	742		
Total	152	320	472	815		
Self-harm						
Edition	2020		2021			
	Train	Test	Train	Test		
Self-harm	41	104	145	152		
Control	299	319	618	1296		
Total	340	423	763	1348		
Gambling						
Edition	2022		2023			
	Train	Test	Train	Test		
Gambling	164	81	245	103		
Control	2184	1998	4182	2071		
Total	2348	2079	4427	2174		

required to make this decision, ERDE is defined as follows:

$$ERDE_o(d,k) = \begin{cases} c_{fp} & \text{if } d \text{ is a False Positive (FP)} \\ c_{fn} & \text{if } d \text{ is a False Negative (FN)} \\ lc_0(k) \cdot c_{tp} & \text{if } d \text{ is a True Positive (TP)} \\ 0 & \text{if } d \text{ is a True Negative (TN)} \end{cases}$$

Here, $lc_0(k)$ is a logistic cost function that increases with the delay k (measured as the number of texts needed to emit the decision). This cost function is defined as follows:

$$lc_0(k) = 1 - \frac{1}{1 + e^{k-o}} \tag{3}$$

ERDE is an error metric, and thus, lower values are indicative of better performance. In our experiments, the delay in decision-making is quantified by the number of user posts (k) processed before the system emits an alert. This design prioritizes early and accurate risk detection.

The parameter o determines the threshold at where the severity of penalties intensifies, reflecting the importance of fast detection. In the eRisk framework, two specific configurations of ERDE are used: $ERDE_5$ and $ERDE_{50}$, where $o = 5$ and $o = 50$, respectively. These configurations represent the number of posts after which the penalization escalates more, with $ERDE_5$ applying stricter penalization for delays beyond five posts. Each classification outcome is associated with a specific cost:

- False positives (FP) and false negatives (FN) incur fixed costs of c_{fp} and c_{fn}, respectively. In the eRisk campaign, c_{fn} was set to 1, while c_{fp} was set to the positive-to-control user ratio in each dataset.
- True positives (TP) incur a time-sensitive cost defined by $lc_0(k)$. Here, $c_{tp} = 1$, reflecting both accuracy and timeliness.
- True negatives (TN) do not incur any cost.

4.2 Implementation Details

4.2.1 Feature Extraction

As previously mentioned, our temporal analysis quantifies word movements (deltas, δ), which rely on having a predetermined vocabulary. This requires aligning the vocabulary derived from the test data with the one corresponding to the training data. This introduces different considerations for each temporal model.

For the DCWE model, the default training process begins with vocabulary filtering, which eliminates stop words and tokens unrecognized by the tokenizer. Following this, we select the top 10,000 words by frequency.[4] DCWE generates one vector per word occurrence, meaning that a single word may have multiple vectorial representations within a chunk. To derive one feature from these representations, we compute the mean of delta values across all vector instances of each word. In contrast, TWEC, the context-independent Word2Vec-based model, generates identical vectors for repeated occurrences of the same word within a chunk. Related to this, TWEC does not involve vocabulary filtering or stop-word removal, as its methodology inherently produces a single representation for each word.

4.2.2 Classification Algorithms

We address each eRisk challenge as a binary classification task, oriented to distinguishing between positive (indicative of a disorder) and control groups of users. For all classifiers, the word shifts within each chunk (deltas) are the only features used for training and prediction. We experimented with two main families of classification methods:

Classic Classification Models For the initial experiments, we explored three classic classification models: support vector machine (SVM), random forest (RF), and

[4] This threshold typically covers all unique words in most datasets, as they usually contain fewer than 10,000 unique tokens.

k-nearest neighbors (KNN). We utilized the implementations of these algorithms available at the scikit-learn library.[5]

- **Support vector machine:** SVM was selected for its strong performance in previous eRisk editions. We performed grid search hyperparameter tuning, optimizing the regularization parameter C (tested across a logarithmic scale from 0.001 to 1000), class weight (adjusted from a 1:1.5 to 1:3.0 ratio with incremental steps of 0.5), and gamma (tested values of 0.1 and 10).
- **Random forest:** RF classifiers were obtained following a grid search to optimize the number of trees (from 100 to 1000 in steps of 100), maximum tree depth (ranging from 10 to 50 in steps of 10), and minimum samples per split (set between 2 and 5). These parameters were tuned to handle the high dimensionality of word-shift features effectively.
- **k-nearest neighbors:** For KNN, we varied the number of neighbors (from 3 to 15), the distance metric (Euclidean or Manhattan), and the weighting strategy (uniform or distance-based). These configurations aimed to identify the optimal settings for the sparse, high-dimensional feature space of temporal deltas.

Neural Networks To further explore the predictive capability of word shifts, we also experimented with some neural network architectures. Two models were employed, both optimized for GPU acceleration:

- **Feed forward neural network (FFNN):** This architecture comprises seven fully connected layers, each separated by ReLU activation functions, with a final sigmoid-activated output layer.
- **Convolutional neural network (CNN):** The CNN model includes three parallel convolutional layers with varying kernel sizes [$input\_dim/2$, $input\_dim/4$, $input\_dim/8$]. Max pooling, feature concatenation, and dropout layers are applied, followed by a sigmoid-activated output layer.

For all models, we conducted hyperparameter tuning using a cross-validation strategy, reserving 30% of the training data for validation. The best-performing parameters, determined by maximizing the F1 score on the validation set, were applied to the test data for final evaluation.

5 Results and Discussion

This section presents the performance obtained by our temporal models on the eRisk datasets (Section 5.1) and their integration with classification algorithms against state-of-the-art baselines (Section 5.2).

[5] https://scikit-learn.org/stable/

5.1 Temporal Models

The primary goal of the initial experiment was to determine which temporal representation model, TWEC or DCWE, performs better across various early detection tasks. To that end, we evaluated both models using the three traditional classifiers, SVM, KNN, and RF.

From the results presented in Table 3, we can see that the DCWE model consistently outperforms TWEC on most datasets and metrics. DCWE achieved superior performance in nearly all comparisons, with the sole exception of the gambling dataset in 2023, where TWEC marginally outperformed DCWE in a single instance. This consistent trend highlights DCWE's stronger ability to capture temporal linguistic variations indicative of early mental health issues. For the ERDE metric, which accounts for both accuracy and delay, DCWE showed lower error rates in nearly all conditions. For example, in the anorexia datasets, DCWE achieved $ERDE_5$ values of 12.90 and 7.78 (using SVM), compared to 21.40 and 8.17 for TWEC. This indicates DCWE's superior ability to make timely and accurate predictions. Notably, the only exception where TWEC outperformed DCWE occurred in the gambling dataset for 2023, where TWEC achieved an $ERDE_5$ value of 2.92 (SVM) compared to 2.99 for DCWE.

By comparatively analyzing the classic classifiers, we can observe that SVM and RF exhibited greater stability across datasets compared to KNN. For example, in the depression datasets (2017 and 2018), both SVM and RF consistently produced higher F1 scores, whereas KNN struggled with notably lower values (e.g., 0.259 for DCWE in 2017). KNN exhibited greater variance in performance and underperformed in most cases, particularly in conditions with smaller datasets. Looking at the overall performance, the F1 scores remained relatively low in most datasets (for example, 0.35 for depression in 2022 and 0.45 for anorexia in 2018), reflecting the inherent difficulty of these tasks. The highest F1 results were found in the gambling 2023 dataset, with an F1 of 0.84 using DCWE and SVM. In order to decrease the computational load of the experimentation and, considering the slight performance disparity between SVMs and RF, we have chosen the DCWE temporal model with SVMs for the following evaluation stages. The comparison and analysis of performance against state-of-the-art methods will be further discussed in the subsequent sections.

5.2 Early Risk Detection

Due to the previous findings and the distinct advantages offered by each temporal model, we have selected the DCWE model for further examination. This subsection evaluates the performance of DCWE with SVM and neural network classifiers, comparing our results against the top-performing models from prior eRisk editions. Specifically, we include in the comparison the best existing system for each metric. We also report the mean performance across all participating systems in the corresponding edition of this shared-data campaign [6]. It is important to highlight that our approach relies exclusively on temporal-based features, whereas many eRisk participants adopted multifaceted approaches. These include incorporating diverse

[6] These average values are shown in the row: *Participants mean.*

Table 3 Comparative performance of DCWE and TWEC models across eRisk datasets using classic classifiers

Depression				
Edition	Model	ERDE		F1
		5	50	
2017	Δ DCWE_SVM	**12.36**	9.28	0.41
	Δ DCWE_RFC	12.44	**7.39**	**0.44**
	Δ DCWE_KNN	12.42	10.90	0.26
	Δ TWEC_SVM	12.56	10.21	0.25
	Δ TWEC_RFC	13.05	11.87	0.19
	Δ TWEC_KNN	12.83	12.20	0.11
2018	Δ DCWE_SVM	9.00	7.10	**0.37**
	Δ DCWE_RFC	**8.72**	**6.03**	0.33
	Δ DCWE_KNN	8.96	8.10	0.26
	Δ TWEC_SVM	9.09	7.84	0.21
	Δ TWEC_RFC	9.63	9.10	0.23
	Δ TWEC_KNN	9.52	9.25	0.11
2022	Δ DCWE_SVM	6.60	4.49	**0.35**
	Δ DCWE_RFC	**6.45**	**3.92**	0.34
	Δ DCWE_KNN	6.54	5.65	0.27
	Δ TWEC_SVM	6.69	5.82	0.17
	Δ TWEC_RFC	6.87	6.28	0.17
	Δ TWEC_KNN	6.76	6.49	0.10
Anorexia				
Edition	Model	ERDE		F1
		5	50	
2018	Δ DCWE_SVM	12.90	**7.56**	**0.45**
	Δ DCWE_RFC	14.69	8.64	0.37
	Δ DCWE_KNN	**11.89**	8.91	0.43
	Δ TWEC_SVM	21.40	15.23	0.22
	Δ TWEC_RFC	17.79	12.30	0.23
	Δ TWEC_KNN	16.32	11.27	0.23
2019	Δ DCWE_SVM	**7.78**	**4.93**	0.57
	Δ DCWE_RFC	8.24	5.62	**0.58**
	Δ DCWE_KNN	7.59	5.96	0.49
	Δ TWEC_SVM	8.17	6.30	0.27
	Δ TWEC_RFC	9.05	7.48	0.21
	Δ TWEC_KNN	8.88	8.28	0.12
Self-harm				
Edition	Model	ERDE		F1
		5	50	

The best outcome for each dataset is bolded. Lower ERDE scores are indicative of superior performance

Table 3 continued

2020	Δ DCWE_SVM	18.23	17.11	0.43
	Δ DCWE_RFC	**15.94**	**12.62**	**0.55**
	Δ DCWE_KNN	24.34	24.31	0.12
	Δ TWEC_SVM	18.79	16.37	0.41
	Δ TWEC_RFC	20.96	14.67	0.42
	Δ TWEC_KNN	23.73	23.20	0.10
2021	Δ DCWE_SVM	9.15	**6.01**	**0.45**
	Δ DCWE_RFC	**8.85**	6.01	0.34
	Δ DCWE_KNN	10.22	10.02	0.13
	Δ TWEC_SVM	9.02	7.21	0.26
	Δ DCWE_RFC	9.68	7.23	0.20
	Δ TWEC_KNN	10.16	9.63	0.11
Gambling				
Edition	Model	ERDE		F1
		5	50	
2022	Δ DCWE_SVM	**3.18**	**1.38**	0.56
	Δ DCWE_RFC	3.29	1.50	**0.63**
	Δ DCWE_KNN	3.24	1.87	0.53
	Δ TWEC_SVM	3.76	3.31	0.13
	Δ TWEC_RFC	3.91	2.29	0.16
	Δ TWEC_KNN	3.89	3.76	0.10
2023	Δ DCWE_SVM	2.99	**0.96**	**0.84**
	Δ DCWE_RFC	3.13	1.30	0.79
	Δ DCWE_KNN	3.25	1.82	0.61
	Δ TWEC_SVM	**2.92**	1.10	0.74
	Δ TWEC_RFC	3.89	2.11	0.77
	Δ TWEC_KNN	3.89	3.76	0.16

The best outcome for each dataset is bolded. Lower ERDE scores are indicative of superior performance

classifiers, exploiting external datasets for pre-training, or employing manually crafted features, such as sentiment analysis or metadata [11, 13, 14, 16, 17, 31].

The results of our comparative analysis are detailed in Table 4. First, we focus on the comparison of the performance across our proposed methods (utilizing SVM, FFNN, and CNN classifiers) against the baselines. A consistent trend emerges across all datasets and conditions: while our methods do not surpass state-of-the-art methods, they generally exceed the average performance of all eRisk participants. For early detection metrics (ERDE), our models achieve competitive results, obtaining the highest $ERDE_5$ in the 2017 depression dataset (left upper block in Table 4). Our methods consistently demonstrate lower ERDE error rates than those achieved by the average participant performance. In terms of F1, our models are closer to the participant mean. For example, in the 2020 self-harm dataset, our best model achieved an F1 score of

Table 4 Comparative performance of DCWE variants and state-of-the-art baselines across eRisk datasets

Depression				
Edition	Model	ERDE		F1
		5	50	
2017	UNSLA [33]	13.66	9.68	0.59
	FHDOA [32]	12.82	9.69	**0.64**
	SS3 [26]	12.70	**7.72**	0.54
	Participants mean	14.67	12.76	0.39
	Δ DCWE_SVM	12.36	9.28	0.41
	Δ DCWE_FFNN	12.27	9.26	0.39
	Δ DCWE_CNN	**12.02**	9.52	0.38
2018	FHDOB [34]	9.50	6.44	**0.64**
	UNSLA [35]	**8.78**	7.39	0.32
	ModComb [50]	9.52	**6.12**	0.51
	Participants mean	10.33	8.24	0.41
	Δ DCWE_SVM	9.00	7.10	0.37
	Δ DCWE_FFNN	8.86	7.22	0.36
	Δ DCWE_CNN	9.63	9.363	0.38
2022	NLPGroup-IISERB [40]	5.50	3.20	**0.71**
	SCIR2 [31]	7.90	**2.60**	0.46
	LauSAn [51]	**3.90**	3.30	0.31
	Participants mean	6.9	4.9	0.30
	Δ DCWE_SVM	6.60	4.49	0.35
	Δ DCWE_FFNN	6.63	4.69	0.33
	Δ DCWE_CNN	6.84	4.65	0.30
Anorexia				
Edition	Model	ERDE		F1
		5	50	
2018	FHDOB [34]	11.98	6.61	**0.85**
	FHDOA [34]	12.15	**5.96**	0.81
	UNSLB [35]	**11.40**	7.82	0.61
	Participants mean	13.31	10.89	0.55
	Δ DCWE_SVM	12.90	7.56	0.45
	ΔDCWE_FFNN	12.42	6.98	0.52
	ΔDCWE_CNN	12.67	7.11	0.51
2019	CLaC4 [37]	6.25	3.43	**0.71**
	UNSL [36]	6.14	**2.97**	0.47
	UNSL [36]	**5.54**	3.92	0.55
	Participants mean	8.60	6.24	0.42
	Δ DCWE_SVM	7.78	4.93	0.57
	ΔDCWE_FFNN	7.21	4.14	0.61
	ΔDCWE_CNN	7.30	4.29	0.59

Table 4 continued

Self-harm				
Edition	Model	ERDE		F1
		5	50	
2020	iLab [52]	25.50	25.50	**0.75**
	iLab [52]	28.70	**7.10**	0.68
	iLab [52]	**13.40**	11.80	0.59
	Participants mean	26.08	17.61	0.47
	Δ DCWE_SVM	18.23	17.11	0.43
	Δ DCWE_FFNN	17.97	15.15	0.51
	Δ DCWE_CNN	18.41	16.01	0.47
2021	UNSL [53]	6.40	3.80	0.62
	UNSL [53]	12.50	3.40	0.49
	UPV-Symanto [54]	**5.90**	5.60	0.38
	Participants mean	10.65	7.33	0.35
	Δ DCWE_SVM	9.15	6.01	0.45
	Δ DCWE_FFNN	9.21	6.511	0.43
	Δ DCWE_CNN	9.17	6.285	0.46
Gambling				
Edition	Model	ERDE		F1
		5	50	
2022	UNSL [55]	2.00	**0.80**	**0.86**
	RELAI [39]	**1.50**	0.90	0.67
	Participants mean	3.37	2.10	0.28
	Δ DCWE_SVM	3.18	1.38	0.56
	Δ DCWE_FFNN	3.09	1.37	0.63
	Δ DCWE_CNN	3.08	1.31	0.60
2023	ELiRF-UPV [56]	**2.60**	**1.00**	**0.93**
	Participants mean	4.76	3.49	0.37
	Δ DCWE_SVM	2.99	0.96	0.84
	Δ DCWE_FFNN	2.96	0.87	0.82
	Δ DCWE_CNN	2.94	0.82	0.81

The best outcome for each dataset is bolded. Lower ERDE scores are indicative of superior performance

0.51, compared to the participant average of 0.47 and the leading method, iLab [52], which achieved 0.75.

Finally, we analyze the performance of our methods when using three distinct classifiers: SVM, FFNN, and CNN. We can observe that their effectiveness is closely matched, varying slightly across different metrics and datasets. For instance, within the 2022 depression dataset, the SVM classifier outperforms the others across all evaluated metrics. Conversely, in the context of the 2019 anorexia dataset, the FFNN classifier emerges as the top performer. Given these observations, no single classifier surpasses the others across all conditions and metrics. This variability underscores the

necessity of selecting classifiers based on the specific characteristics of each dataset and the metrics of interest.

Performance and Future Potential We attribute the observed performance gaps between our methods and the state-of-the-art systems to the different strategies employed. Most state-of-the-art approaches leverage an ensemble of models and techniques, integrate multiple features (e.g., textual, metadata, or temporal features), or utilize additional external datasets tailored to the target domains [11, 13, 14, 16, 17, 31]. By contrast, our approach focuses only on temporal word movement features. This simple and scalable design has substantial potential for further optimization. For instance, our method could be augmented by integrating complementary features, such as sentiment analysis, user metadata, or textual patterns.

Moreover, the inherent difficulty of these tasks cannot be overstated, as reflected in the performance ranges observed. This difficulty underscores the challenges posed by early risk detection. Our results point to potential avenues for improvement, such as concentrating on domain-specific vocabularies (instead of general vocabularies) or integrating our temporal embeddings with complementary methods from prior literature. These enhancements could broaden the potential applicability of our methods, which could become robust components within a broader system.

6 Keyword Evolution

To observe the dynamics of word usage and its significance from an analytical perspective, this section presents an analysis aimed at identifying: (1) words that influence the performance of our methods (i.e., mental-associated keywords) and (2) the role these keywords play in differentiating between positive and control users. We employed a point-biserial correlation analysis across our experiments, correlating the target labels (positive and control) with the features representing word movements. This statistical approach, a variant of the Pearson correlation [57], enables the examination of relationships between continuous variables (word movements) and a binary variable (user classification). This analysis focuses on two disorders: gambling, where our methods achieved the best performance, and depression, where outcomes were less favorable.

In our analysis of the gambling 2023 dataset, we analyze the dynamics of word usage and its correlation with classification outcomes. Figure 4 displays the top ten words, ordered by their correlation weight with the classification labels across both training and testing datasets. These terms, closely related to gambling, highlight their relevance in tracking linguistic shifts that may signal gambling-related issues. Based on these results, we designed plots that examine the trends of the five most correlated words with gambling disorder: “gambling,” “gamble,” “money,” “lost,” and “addiction.”

The first plot (Fig. 5) visualizes the evolution of these word shifts among positive and control groups over the ten temporal chunks, with the y-axis indicating the average word movement and the x-axis representing the temporal chunks. The average movement is calculated as the sum of word movements divided by the number of occurrences in each chunk. The plot reveals a distinct pattern among groups: the positive group shows a greater fluctuation in the use of these key gambling-related terms.

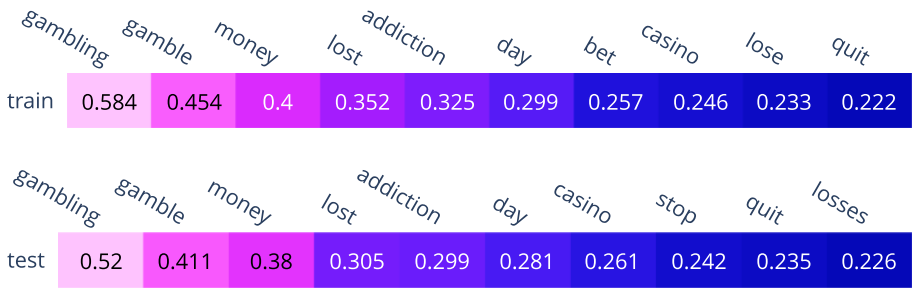


Fig. 4 Top 10 correlated words with the positive label in the gambling datasets

Complementing this, Fig. 6 shows the percentage of users mentioning each keyword across the temporal chunks. This second plot underscores a contrast in word usage between groups, where the control group uses these words less frequently, with no single word exceeding 20% usage. The different usage pattern between the positive gambling users and the control group further illustrates the potential of specific word usage trends in identifying gambling-related concerns.

We conducted a similar analysis on the 2022 depression dataset, where our models exhibited modest performance. Depression is a broader and more complex disorder, which often leads to discussions that span various aspects of life and emotions [8]. This contrasts with gambling, a more focused topic, where specific word usage is more directly indicative of the disorder. This distinction means that using words for monitoring depression may become more challenging. Figure 7 presents the top ten words most correlated with the depression group. While these top words are relevant to depression, the correlation weights are considerably lower than those in the gambling dataset. For example, the value for the top word in the depression training dataset ("feel") has a correlation value of 0.253, markedly lower than the top word in the gambling dataset ("gambling"), which reaches 0.584.

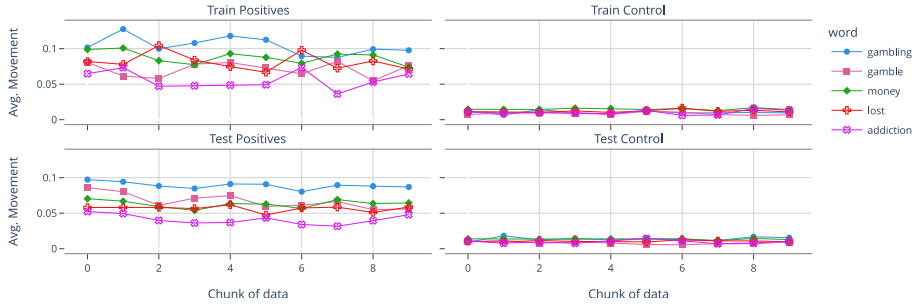


Fig. 5 Average word movements over the temporal chunks and the top 5 correlated words in the gambling 2023 dataset

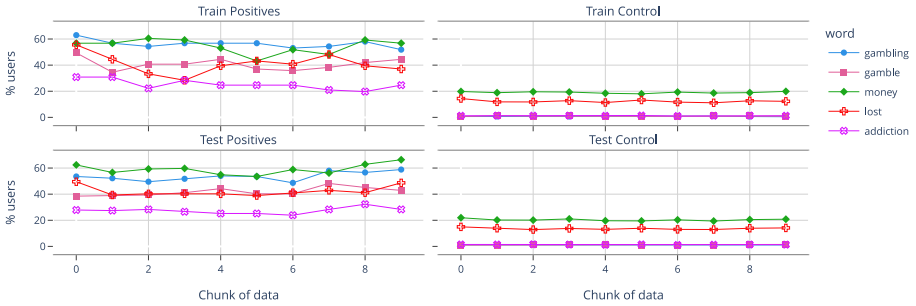


Fig. 6 Percentage of users using the top 5 correlated works over the temporal chunks in the gambling 2023 dataset

Further illustrating this point, Fig. 8 displays the word movements for the top five correlated words in the depression dataset. Again, we can see differences in word shifts between depressed and control groups. Additionally, Fig. 9 examines the percentage of users mentioning these words. While general terms like "feel" and "life" predominate among control users, specific terms such as "anxiety" and "suicidal" are less common but more frequently mentioned by the positive group. This analysis underscores the intricate relationship between language and mental health, emphasizing the need for advanced, context-aware models capable of capturing subtle linguistic cues specific to depression.

7 Exploring Alternative Distance Metrics

In this section, we explore alternative distance metrics to estimate word movements within our framework of temporal embeddings. These metrics replace the cosine distance metric used to compute the deltas (Section 3). By exploring alternative distance functions, we aim to assess their influence on the early detection of various mental health disorders.

To compute word deltas, we calculate the distance between the contextual embedding of a word in a given chunk ($C^{t_j}(w)$) and its reference embedding ($R(w)$). In

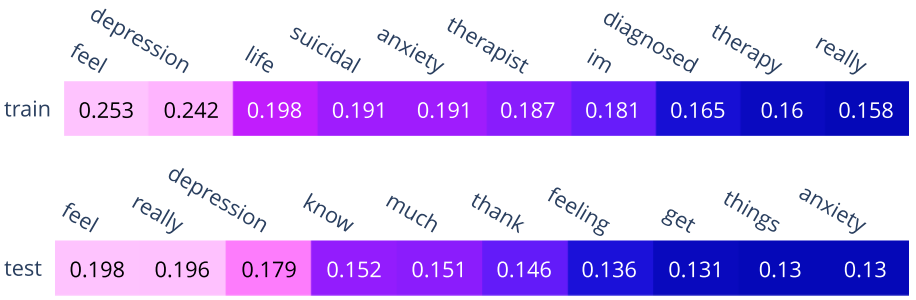


Fig. 7 Top 10 correlated words with the positive label in the depression 2022 dataset

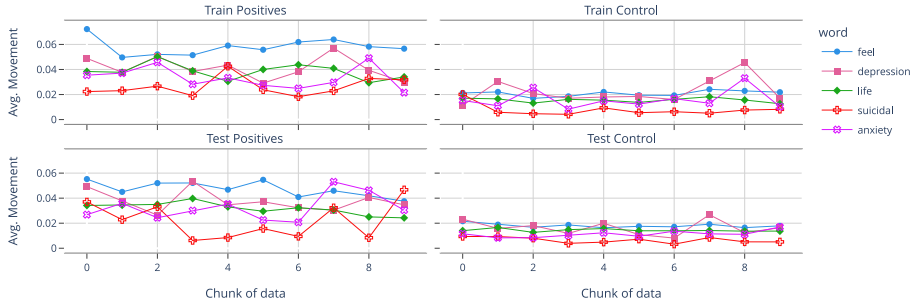


Fig. 8 Average word movements over the temporal chunks and the top 5 correlated words in the depression 2022 dataset

the previously reported experiments, the cosine distance between embeddings was employed. However, alternative metrics, which capture different geometrical properties, may better reflect the semantic shifts. Here, we examine three additional distance metrics[7]:

- **Euclidean distance** (L_2): Measures the straight-line distance in the embedding space. It calculates the magnitude of vector differences as follows:

$$\delta_w^{t_j} = \begin{cases} \sqrt{\sum_{i=1}^{n} (C^{t_j}(w)_i - R(w)_i)^2}, & \text{if } w \text{ occurs in chunk } t_j, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

- **Manhattan distance** (L_1): Computes the sum of absolute differences between vector elements, calculating coordinate-wise discrepancies:

$$\delta_w^{t_j} = \begin{cases} \sum_{i=1}^{n} |C^{t_j}(w)_i - R(w)_i|, & \text{if } w \text{ occurs in chunk } t_j, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

- **Chebyshev distance:** It is oriented to extract the largest difference along any single dimension of the embedding vectors:

$$\delta_w^{t_j} = \begin{cases} \max_{i=1,\ldots,n} |C^{t_j}(w)_i - R(w)_i|, & \text{if } w \text{ occurs in chunk } t_j, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Table 5 presents the comparative performance of these distance metrics across the eRisk datasets, using DCWE as the temporal embedding model and SVM as the classifier. The ERDE and F1 results indicate that cosine distance consistently outperforms other distance metrics across all datasets. This superiority can be attributed to cosine's focus on directional relationships, which aligns closely with the semantic properties of word embeddings. Cosine distance is widely recognized as an effective metric for comparing vector embeddings, particularly in tasks like semantic search [58]. The Euclidean and Manhattan distances also demonstrate competitive results, with minor

[7] The subscript i in a vector v, denoted as v_i, refers to the i-th element in v.

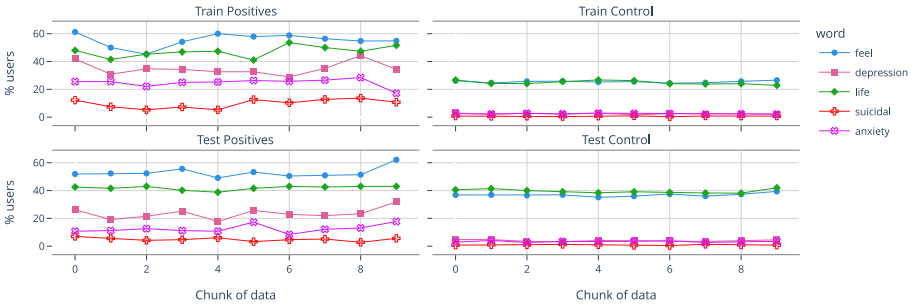


Fig. 9 Percentage of users using the top 5 correlated works over the temporal chunks in the depression 2022 dataset

Table 5 Comparative performance of different distance metrics using DCWE with SVM classifiers

Depression				
Edition	Distance	ERDE		F1
		5	50	
2017	Cosine	**12.36**	**9.28**	**0.41**
	Chebyshev	12.98	10.20	0.36
	Manhattan	13.01	10.22	0.35
	Euclidean	12.95	10.04	0.37
2018	Cosine	**9.00**	**7.10**	**0.37**
	Chebyshev	9.56	8.14	0.31
	Manhattan	9.62	8.26	0.28
	Euclidean	9.57	8.16	0.32
2022	Cosine	**6.60**	**4.49**	0.35
	Chebyshev	6.91	5.41	0.32
	Manhattan	6.88	5.37	0.34
	Euclidean	6.88	5.27	**0.37**
Anorexia				
Edition	Distance	ERDE		F1
		5	50	
2018	Cosine	**12.90**	**7.56**	**0.45**
	Chebyshev	12.97	7.68	0.44
	Manhattan	12.91	7.66	0.44
	Euclidean	14.09	8.53	0.39
2019	Cosine	**7.78**	**4.93**	**0.57**
	Chebyshev	8.87	6.67	0.44
	Manhattan	8.83	6.51	0.43
	Euclidean	8.87	6.59	0.45
Self-harm				
Edition	Distance	ERDE		F1
		5	50	

The best outcome for each dataset is bolded. Lower ERDE scores are indicative of superior performance

Table 5 continued

2020	Cosine	**18.23**	**17.11**	**0.43**
	Chebyshev	23.86	22.87	0.13
	Manhattan	23.41	22.26	0.17
	Euclidean	23.65	22.78	0.13
2021	Cosine	**9.15**	6.01	0.45
	Chebyshev	9.82	6.86	0.44
	Manhattan	9.18	6.23	0.44
	Euclidean	9.23	**5.92**	**0.46**
Gambling				
Edition	Distance	ERDE		F1
		5	50	
2022	Cosine	**3.18**	**1.38**	**0.56**
	Chebyshev	3.77	2.13	0.54
	Manhattan	3.75	2.08	0.49
	Euclidean	3.75	2.07	0.50
2023	Cosine	**2.99**	**0.96**	**0.84**
	Chebyshev	3.43	1.33	0.77
	Manhattan	3.34	1.20	0.80
	Euclidean	3.35	1.21	0.79

The best outcome for each dataset is bolded. Lower ERDE scores are indicative of superior performance

differences between them. For instance, Manhattan distance achieves an F1 score of 0.44 in the 2019 anorexia dataset, closely aligning with Euclidean's 0.45. On the other hand, Chebyshev's distance generally underperforms compared to the other metrics.

Notably, there are specific cases where these alternative metrics achieve marginally better results than those achieved by cosine distance. For example, in the 2021 self-harm dataset, the Euclidean metric achieves a slightly lower $ERDE_{50}$ score than cosine (5.92 vs. 6.01). Nevertheless, cosine remains the most robust metric overall, particularly for the gambling datasets, where it achieves consistently superior results. These findings confirm that cosine distance is a robust choice for estimating word movements in the DCWE framework. However, alternative metrics may offer complementary perspectives for specific datasets or applications. For instance, Manhattan distance, which emphasizes element-wise differences, may be useful for tasks where small shifts in individual dimensions are critical. Future research could explore hybrid approaches that combine multiple metrics to capture diverse linguistic nuances. Additionally, domain-specific adaptations, such as weighting certain words in the vocabulary, could further improve the effectiveness of these metrics.

8 Conclusions and Future Work

In this paper, we explored the role of language evolution as an early indicator of mental health issues. To that end, we utilized social media data from all the eRisk collections,

analyzing datasets corresponding to four primary disorders: depression, anorexia, self-harm, and pathological gambling. These datasets represent a diverse range of mental health challenges, containing over 1000 positive cases and 12,000 control users. A main hypothesis guides our work: the progression of a mental health condition is often accompanied by changes in language use. To explore this aspect, we proposed a framework to capture and quantify word usage evolution over time. This framework leverages temporal word embeddings, utilizing static (Word2Vec) and contextualized (BERT) models to track word movements over time. Our classifiers relied only on these word shift patterns as features, without requiring disorder-specific features or complex decision-making algorithms. Our approach demonstrated competitive early detection capabilities across all evaluated datasets, highlighting the effectiveness of temporal word movement features as mental health indicators. Moreover, we included a keyword evolution analysis to visualize the most indicative words for each disorder, providing deeper insights into how specific terms and word changes can reflect users' mental states on social media. The flexibility of our temporal-aware framework paves the way for numerous avenues of future work. To improve our performance, we plan to refine our models by incorporating disorder-specific lexica or integrating additional features such as sentiment analysis, user metadata, or multimodal data. Additionally, we plan to extend our analysis to cover a broader range of mental health conditions, improving the robustness and applicability of our word evolution framework. Another promising direction is exploring transfer learning or cross-domain approaches, where models trained on one dataset are evaluated on test sets from other years or disorders. Such experiments could assess the generalizability of our methods and uncover potential for wider application in diverse mental health scenarios.

Acknowledgements The second and third authors thank the financial support supplied from projects: PLEC2021-007662 (MCIN/AEI/10.13039/501100011033 Ministerio de Ciencia e Innovación, European Union NextGenerationEU/PRTR) and PID2022-137061OB-C21 (MCIN/AEI/10.13039/501100011033/, Ministerio de Ciencia e Innovación, ERDF A way of making Europe, by the European Union); Consellería de Educación, Universidade e Formación Profesional, Spain (accreditations 2019-2022 ED431G/01 and GPC ED431B 2022/33) and the European Regional Development Fund, which acknowledges the CITIC Research Center. The first and fourth authors thank the financial support supplied by the Agencia Estatal de Investigación (Spain) (PID2022-137061OB-C22; PLEC2021-007662 MCIN/AEI/10.13039/501100011033, Plan de Recuperación, Transformación y Resiliencia, Unión Europea-Next Generation EU), Consellería de Cultura, Educación, Formación Profesional e Universidades (Centro de investigación de Galicia accreditation 2024-2027 ED431G-2023/04 and Reference Competitive Group accreditation 2022-2025, ED431C 2022/19) and the European Union (European Regional Development Fund - ERDF). The fourth author thanks the support obtained from project SUBV23/00002 (Ministerio de Consumo, Subdirección General de Regulación del Juego).

Author Contribution M.C.: conceptualization, implementation, investigation, writing, and supervision. A.P.: conceptualization, implementation, investigation, writing, and supervision. J. P: conceptualization, investigation, and supervision. D.E.L: conceptualization, investigation, and supervision.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Data Availability No datasets were generated or analyzed during the current study.

Declarations

Conflict of Interest The authors declare no competing interests.

References

1. Oyebode O, Ndulue C, Mulchandani D, Suruliraj B, Adib A, Orji FA, Milios E, Matwin S, Orji R (2022) COVID-19 pandemic: identifying key issues using social media and natural language processing. J Healthcare Inf Res 6(2):174–207
2. World Health Organization: depressive disorder (depression) (2023). https://www.who.int/news-room/fact-sheets/detail/depression, Last accessed on 2024-01-09
3. Halfin A (2007) Depression: the benefits of early and appropriate treatment. American J Manage Care 13(4):92
4. Picardi A, Lega I, Tarsitani L, Caredda M, Matteucci G, Zerella M, Miglio R, Gigantesco A, Cerbo M, Gaddini A et al (2016) A randomised controlled trial of the effectiveness of a program for early detection and treatment of depression in primary care. J Affect Disorders 198:96–101
5. Pennebaker JW, Boyd RL, Jordan K, Blackburn K (2015) The development and psychometric properties of LIWC2015. Tech Report
6. Ríssola EA, Losada DE, Crestani F (2021) A survey of computational methods for online mental state assessment on social media. ACM Trans Comput Healthcare 2(2)
7. Aldkheel A, Zhou L (2024) Depression detection on social media: a classification framework and research challenges and opportunities. J Healthcare Inf Res 8(1):88–120
8. De Choudhury M, Counts S, Horvitz E (2013) Social media as a measurement tool of depression in populations. In: Proceedings of the 5th Annual ACM Web Science Conference, pp 47–56
9. Masino AJ, Forsyth D, Fiks AG (2018) Detecting adverse drug reactions on twitter with convolutional neural networks and word embedding features. J Healthcare Inf Res 2:25–43
10. Zakkar MA, Lizotte DJ (2021) Analyzing patient stories on social media using text analytics. J Healthcare Inf Res 5(4):382–400
11. Parapar J, Martín-Rodilla P, Losada DE, Crestani F (2023) Overview of eRisk 2023: early risk prediction on the Internet. In: International Conference of the Cross-Language Evaluation Forum for European Languages, pp 294–315 . Springer
12. Chim J, Tsakalidis A, Gkoumas D, Atzil-Slonim D, Ophir Y, Zirikly A, Resnik P, Liakata M (2024) Overview of the CLPsych 2024 shared task: leveraging large language models to identify evidence of suicidality risk in online posts. In: Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024), pp 177–190
13. Losada DE, Crestani F, Parapar J (2017) eRisk 2017: CLEF lab on early risk prediction on the Internet: experimental foundations. In: International Conference of the Cross-Language Evaluation Forum for European Languages, pp 346–360. Springer
14. Losada DE, Crestani F, Parapar J (2018) Overview of eRisk: early risk prediction on the internet. In: International Conference of the Cross-language Evaluation Forum for European Languages, pp 343–361 . Springer
15. Losada DE, Crestani F, Parapar J (2019) Overview of eRisk 2019 early risk prediction on the Internet. In: International Conference of the Cross-Language Evaluation Forum for European Languages, pp 340–357. Springer
16. Losada DE, Crestani F, Parapar J (2020) eRisk 2020: self-harm and depression challenges. In: European Conference on Information Retrieval, pp 557–563 . Springer
17. Parapar J, Martín-Rodilla P, Losada DE, Crestani F (2021) Overview of eRisk 2021: early risk prediction on the Internet. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 12th

International Conference of the CLEF Association, CLEF 2021, Virtual Event, 21-24 September, 2021, Proceedings, pp 324–344
18. Hamilton WL, Leskovec J, Jurafsky D (2016) Diachronic word embeddings reveal statistical laws of semantic change. arXiv:1605.09096
19. Yao Z, Sun Y, Ding W, Rao N, Xiong H (2018) Dynamic word embeddings for evolving semantic discovery, pp 673–681
20. Rudolph M, Blei D (2018) Dynamic embeddings for language evolution. In: Proceedings of the 2018 World Wide Web Conference. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE . WWW '18, pp 1003–1011
21. Di Carlo V, Bianchi F, Palmonari M (2019) Training temporal word embeddings with a compass. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 33, pp 6326–6334
22. Hofmann V, Pierrehumbert J, Schütze H (2021) Dynamic contextualized word embeddings. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics
23. Devlin J, Chang M, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, 2-7 June, 2019, Volume 1 (Long and Short Papers), pp 4171–4186. Association for Computational Linguistics, Minneapolis, MN, USA
24. Couto M, Pérez A, Parapar J (2022) Temporal word embeddings for early detection of signs of depression. In: Proceedings of the 2nd Joint Conference of the Information Retrieval Communities in Europe (CIRCLE 2022), 4-7 July, 2022. CEUR Workshop Proceedings, vol 3178. CEUR-WS.org, Samatan, Gers, France
25. Pérez A, Parapar J, Barreiro Á, Lopez-Larrosa S (2023) BDI-Sen: a sentence dataset for clinical symptoms of depression. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 2996–3006
26. Burdisso SG, Errecalde M, Montes-y-Gómez M (2019) A text classification framework for simple and effective early depression detection over social media streams. Exp Syst Appl 133:182–197
27. De Choudhury M, De S (2014) Mental health discourse on reddit: self-disclosure, social support, and anonymity. In: Eighth International AAAI Conference on Weblogs and Social Media
28. Ortega-Mendoza RM, Hernández-Farías DI, Montes-y-Gómez M, Villaseñor-Pineda L (2022) Revealing traces of depression through personal statements analysis in social media. Art Intell Med 123:102202
29. Cohan A, Desmet B, Yates A, Soldaini L, MacAvaney S, Goharian N (2018) SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. arXiv:1806.05258
30. Losada DE, Crestani F (2016) A test collection for research on depression and language use. in proceedings conference and labs of the evaluation forum clef 2016, evora, portugal., 28–39 . Springer
31. Parapar J, Martín-Rodilla P, Losada DE, Crestani F (2022) eRisk 2022: pathological gambling, depression, and eating disorder challenges. In: European Conference on Information Retrieval, pp 436–442 . Springer
32. Trotzek M, Koitka S, Friedrich CM (2017) Linguistic metadata augmented classifiers at the CLEF 2017 task for early detection of depression. In: CLEF (Working Notes)
33. Villegas MP, Funez DG, Ucelay MJG, Cagnina LC, Errecalde ML (2017) LIDIC-UNSL's participation at eRisk 2017: pilot task on early detection of depression. In: CLEF (Working Notes)
34. Trotzek M, Koitka S, Friedrich CM (2018) Word embeddings and linguistic metadata at the CLEF 2018 tasks for early detection of depression and anorexia. In: CLEF (Working Notes)
35. Funez DG, Ucelay MJG, Villegas MP, Burdisso S, Cagnina LC, Montes-y-Gómez M, Errecalde M (2018) UNSL's participation at eRisk 2018 lab. In: CLEF (Working Notes)
36. Burdisso SG, Errecalde M, Montes-y-Gómez M (2019) UNSL at eRisk 2019: a unified approach for anorexia, self-harm and depression detection in social media. In: CLEF (Working Notes)
37. Mohammadi E, Amini H, Kosseim L (2019) Quick and (maybe not so) easy detection of anorexia in social media posts. In: CLEF (Working Notes)
38. Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota

39. Stalder S, Zankov E (2022) ZHAW at eRisk 2022: predicting signs of pathological gambling-glove for snowy days. In: CLEF (Working Notes), pp 987–994
40. Talha A, Basu T (2023) A natural language processing based risk prediction framework for pathological gambling. Working Notes of CLEF, 18–21
41. Trotzek M, Koitka S, Friedrich CM (2020) Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. IEEE Trans Knowl Data Eng 32(3):588–601
42. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv:1301.3781
43. Pennington J, Socher R, Manning CD (2014) GloVe: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 1532–1543
44. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ (2020) Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res 21:140–114067
45. Kim Y, Chiu Y-I, Hanaki K, Hegde D, Petrov S (2014) Temporal analysis of language through neural language models. In: Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, pp 61–65. Association for Computational Linguistics, Baltimore, MD, USA
46. Giulianelli M, Del Tredici M, Fernández R (2020) Analysing lexical semantic change with contextualised word representations. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp 3960–3973. Association for Computational Linguistics, Online
47. Hu R, Li S, Liang S (2019) Diachronic sense modeling with deep contextualized word embeddings: an ecological view. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp 3899–3908. Association for Computational Linguistics, Florence, Italy
48. Schlechtweg D, McGillivray B, Hengchen S, Dubossarsky H, Tahmasebi N (2020) SemEval-2020 task 1: unsupervised lexical semantic change detection. arXiv:2007.11464
49. Kulkarni V, Al-Rfou R, Perozzi B, Skiena S (2015) Statistically significant detection of linguistic change. In: Proceedings of the 24th International Conference on World Wide Web, pp 625–635
50. Ramiandrisoa F, Mothe J (2020) Early detection of depression and anorexia from social media: a machine learning approach. In: Circle 2020, vol 2621
51. Säuberli A, Cho S, Stahlhut L (2022) Lausan at eRisk 2022: simply and effectively optimizing text classification for early detection. In: CLEF (Working Notes), pp 995–1004
52. Martínez-Castaño R, Htait A, Azzopardi L, Moshfeghi Y (2020) Early risk detection of self-harm and depression severity using BERT-based transformers: ilab at clef erisk 2020
53. Loyola JM, Burdisso S, Thompson H, Cagnina LC, Errecalde M (2011) UNSL at eRisk 2021: a comparison of three early alert policies for early risk detection. In: CLEF (Working Notes), pp 992–1021
54. Basile A, Chinea-Rios M, Uban A-S, Müller T, Rössler L, Yenikent S, Chulvi-Ferriols MA, Rosso P, Franco-Salvador M (2021) UPV-Symanto at eRisk 2021: mental health author profiling for early risk prediction on the internet. In: Proceedings of the Working Notes of CLEF 2021, Conference and Labs of the Evaluation Forum, Bucharest, Romania, 21st to 24th September, 2021, pp 908–927. CEUR
55. Saravani SHH, Normand L, Maupomé D, Rancourt F, Soulas T, Besharati S, Normand A, Mosser S, Meurs M-J (2022) Measuring the severity of the signs of eating disorders using similarity-based models. In: CLEF (Working Notes), pp 936–946
56. Molina A, Huang X, Hurtado L-F, Pla F (2023) ELIRF-UPV at eRisk 2023: early detection of pathological gambling using SVM
57. Sedgwick P (2012) Pearson's correlation coefficient. Bmj 345
58. Reimers N, Gurevych I (2019) Sentence-BERT: sentence embeddings using Siamese BERT-networks. In: Inui K, Jiang J, Ng V, Wan X (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp 3982–3992. Association for Computational Linguistics, Hong Kong, China

4.4 Publication V: LabChain: Enabling Reproducible and Modular Scientific Experiments in Python

Authors: Manuel Couto, Javier Parapar, David E. Losada
Published in: SoftwareX [18]
Keywords: Scientific Workflows, Pipeline Architecture, Hash-based Caching, Reproducible Research, Software Engineering Practices

SoftwareX 33 (2026) 102543

Contents lists available at ScienceDirect

SoftwareX

journal homepage: www.elsevier.com/locate/softx

LabChain: Enabling reproducible and modular scientific experiments in Python

Manuel Couto [a,*], Javier Parapar [b], David E. Losada [a]

[a] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain
[b] Centro de Investigación en Tecnologías de la Información y las Comunicaciones (CITIC), Universidade de A Coruña, A Coruña, Spain

ARTICLE INFO

Keywords:
Scientific workflows
Pipeline architecture
Hash-based caching
Reproducible research
Software engineering practices

ABSTRACT

Python's flexibility accelerates research prototyping but frequently results in unmaintainable code and duplicated computational effort. The absence of software engineering practices in academic development leads to fragile experiments where even minor modifications require rerunning expensive computations from scratch. LabChain addresses this through a pipeline-and-filter architecture with hash-based caching that automatically identifies and reuses intermediate results. When evaluating multiple classifiers on the same embeddings, the framework computes embeddings once—regardless of how many classifiers are tested. This automatic reuse extends across research teams: if another researcher applies different models to the same preprocessed data, LabChain detects existing results and eliminates redundant computation. Beyond efficiency, the framework's modular structure reduces technical debt that obscures experimental logic. Pipelines serialize to JSON for reproducibility and distributed execution across computational clusters. A mental health detection case study demonstrates dual impact: computational savings exceeding 12 hours per task with reduced CO_2 emissions, alongside substantial scientific improvements—performance gains up to 192.3% in some tasks. These improvements emerged from clearer experimental organization that exposed a critical preprocessing bug hidden in the original monolithic implementation. LabChain proves that software engineering discipline amplifies scientific discovery.

Metadata

Code metadata		
Nr.	Code metadata description	Metadata
C1	Current code version	v1.2.1
C2	Permanent link to code/repository	https://github.com/manucouto1/LabChain
C3	Permanent link to Reproducible Capsule	https://manucouto1.github.io/LabChain/examples/notebooks/optuna_optimizer_with_data_splitter_kfold/
C4	Legal Code License	AGPL-3.0
C5	Code versioning system used	Git
C6	Software languages, tools, and services	Python
C7	Compilation requirements, operating environments & dependencies	Python 3.11, typeguard 4.4.1, multimethod 1.12, pyspark 3.5.3, fastapi 0.115.5, pandas 2.2.3, torch 2.6.0, scipy 1.13.1, rich 13.9.4, boto3 1.35.73, scikit-learn 1.5.2, cloudpickle 3.1.0, tqdm 4.67.1, nltk 3.9.1, transformers 4.46.3, gensim 4.3.3, wandb 0.18.7, optuna 4.2.1, sentence-transformers 4.0.2
C8	Link to developer documentation/manual	https://manucouto1.github.io/LabChain/
C9	Support email for questions	manuel.couto.pintos@usc.es

* Corresponding author.
Email address: manuel.couto1@udc.es, manuel.couto.pintos@usc.es (M. Couto).

https://doi.org/10.1016/j.softx.2026.102543
Received 20 November 2025; Received in revised form 22 January 2026; Accepted 25 January 2026
Available online 5 February 2026

1. Motivation and significance

In the field of data science and machine learning, good practices and design principles play a key role in improving the efficiency, reproducibility, and scalability of experiments. Despite progress in standardizing tools for data processing and model building, the research community still faces challenges related to duplicated efforts, poor integration between different code components, and inefficient management of computational resources.

There are several open-source frameworks that attempt to address some of these issues, including scikit-learn, Kedro, Snakemake, PyCaret, and Luigi, among others. Each of these frameworks aims to facilitate the construction and execution of machine learning and data processing pipelines. For example, scikit-learn [1] is well-known for its support for building reusable models through pipelines and its broad collection of transformers and estimators, which simplify model reuse and validation. Kedro [2], in turn, specializes in building modular projects that enable large-scale workflow management and reproducibility, while Snakemake [3] provides a unified, transparent, and adaptable way to represent complete data analysis workflows, from raw data processing to final result visualization. PyCaret [4] offers a high-level solution that automates the modeling process without sacrificing modularity. Luigi [5], for its part, focuses on orchestrating complex workflows, particularly in distributed processing contexts.

Despite the advantages these frameworks offer [6], there are still important areas for improvement. Many of them do not efficiently address code duplication, the management of intermediate data, or the reuse of computational components across research teams working on similar tasks. For instance, in research settings involving multi-stage data processing pipelines, some steps are often computationally demanding and redundantly executed across various experiments. A typical example is the generation of embeddings using a pretrained neural model over a large textual corpus (e.g., applying BERT to obtain vectorial representations of documents). When embeddings are not shared across tasks, each experiment redundantly recomputes them, resulting in unnecessary computational overhead and substantial environmental impact.

Let us imagine the following scenario: two research teams are working on experiments that define different pipelines. These pipelines consist of various processing stages that we will refer to as filters. In the case of the first team, the pipeline is composed of filters A, B, and C, where steps A and B are extremely computationally expensive. Meanwhile, the second team is conducting a similar study, but their pipeline includes filters A, B, and D, among other possible variations.

Both teams could greatly benefit from a system that allows them to share filters (A, B) and avoid redundant computations. Our framework, LabChain, makes it possible to structure experiments in such a way that the computation of steps A and B is performed only once. LabChain handles unique identifiers that allow intermediate results to be stored and reused, thus eliminating the need to re-run the same transformations. This caching mechanism is illustrated in Fig. 1.

One of LabChain's most notable innovations is its ability to serialize and deserialize entire pipelines in JSON format. This allows packaging experiments as JSON files for transmission over standard protocols such as REST, GraphQL, gRPC, or event-based systems (Kafka, MQTT). This facilitates configuration exchange between users and enables remote compute nodes to receive serialized pipelines, deserialize them, and execute autonomously. As a result, it is possible to build distributed processing architectures where different nodes test experimental variations without manual intervention, fostering reproducibility and scalability essential to computational science.

1.1. Comparative analysis with related systems

LabChain occupies a unique niche among workflow frameworks by prioritizing automatic inter-team cache reuse and executable configuration over traditional workflow orchestration. While frameworks like scikit-learn, Kedro, Snakemake, and Luigi provide valuable pipeline abstractions, they do not address fundamental inefficiencies in collaborative research, such as the redundant computation of identical transformations across teams and the separation between workflow definition and execution logic.

The inter-team reuse problem. Consider a research group where multiple members work on related experiments using the same preprocessing steps (e.g., BERT embeddings over a corpus). In traditional frameworks, each researcher either recomputes transformations (scikit-learn), manually coordinates file sharing (Kedro/Luigi), or references explicit file paths in workflow rules (Snakemake). LabChain utilizes content-addressable storage, which enables identical configurations and input data to automatically produce cache hits across researchers without requiring manual coordination. This works across local and cloud storage backends (S3), enabling seamless collaboration.

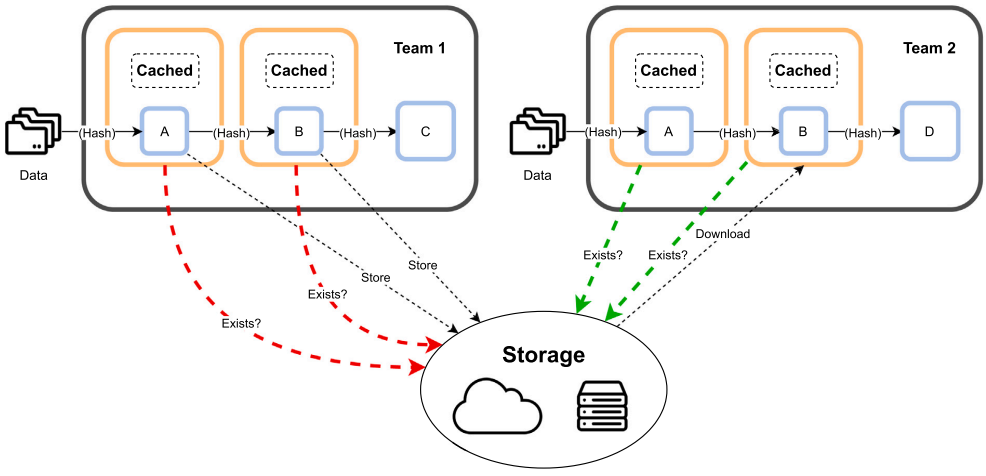


Fig. 1. LabChain: caching mechanism.

Table 1
Comparison of LabChain with existing workflow frameworks.

Feature	LabChain	scikit-learn	Kedro	Snakemake
Pipeline serialization	JSON (executable)	Pickle	YAML + code	YAML (rules)
Automatic caching	Hash-based	None	Manual	Timestamps
Inter-team reuse	Automatic	No	Manual	Manual
State management	Filter-internal	Pipeline objects	Catalog flow	File artifacts
Workflow definition	Code (filters)	Code	YAML DAG	YAML rules
Cloud storage	Native (S3)	No	Plugin	Limited
Setup overhead	Minimal	None	High	Medium
Target use case	Iterative research	Model building	Production	Bioinformatics

Executable configuration vs. declarative workflows. A key architectural distinction lies in how LabChain unifies configuration and execution logic. Kedro or Snakemake separate workflow definition (YAML/Python DAGs) from execution code. State (fitted models) flows through the pipeline alongside data, requiring explicit input/output declarations. In LabChain, configuration *is* executable code. A JSON-serialized pipeline can be directly instantiated and executed. Data flow is implicit in the filter sequence. State lives within filter instances as private attributes, not as pipeline data.

This design provides two practical advantages: (1) each JSON dump represents a complete, reproducible experiment that serves as an experimental backlog of explored configurations, and (2) researchers think in terms of sequential transformations rather than dependency graphs, with state management encapsulated within components.

Conceptual contribution. While individual components (pipelines, caching, dependency injection) are not novel, LabChain's contribution lies in their integration for collaborative experimental research with automatic result reuse. This differs from production workflow orchestration (Kedro, Luigi), bioinformatics workflows (Snakemake), and model building (scikit-learn), each of which targets distinct use cases with varying trade-offs. Table 1 provides a comparative analysis between LabChain and other existing workflow frameworks.

2. Software description

LabChain is built on established design principles and software patterns. Its architecture encapsulates each processing step as a *filter*, which can be combined into serializable pipelines. This approach facilitates reuse, composition, and distributed execution. The framework employs a container-based dependency injection system alongside a hash-based storage strategy reminiscent of content-addressable storage systems.

2.1. Filter architecture and serialization

LabChain follows the **Pipes and Filters** pattern. Each filter, derived from `BaseFilter`, encapsulates a specific data transformation. To overcome the limitations of standard JSON serialization when handling complex machine learning objects (such as tokenizers, vectorizers, or pretrained models), LabChain enforces a clear separation of attributes:

- **Public Attributes (Configuration):** These are the filter's configuration parameters that define its behavior. LabChain serializes these attributes and uses them as constructor arguments during filter reconstruction (passed as `**kwargs` to `__init__`). Consequently, all public attributes must be declared as `__init__` parameters. Only basic serializable types (strings, integers, floats, booleans, lists, dictionaries) are supported.
- **Private Attributes (Non-configuration State):** Private attributes (prefixed with an underscore) represent internal implementation details rather than configuration parameters. These include execution state, utility objects, and non-serializable components that do not define the filter's identity.

This design allows `item_dump()` to produce portable JSON representations of a pipeline "recipe" without the overhead of serializing arbitrary Python objects.

2.2. Cache-key specification and validation

Reproducibility in LabChain relies on an explicit and deterministic cache-key definition. Cache keys are computed as cryptographic hashes derived exclusively from user-visible and serialized information, ensuring that identical experimental configurations yield identical cache identifiers.

Deterministic hashing modes. LabChain implements a dual-mode hashing system adapted to filter characteristics:

- **Non-trainable filters (initialization-based):** For static transformations, the hash is computed at instantiation from the class name and all public attribute values.
- **Trainable filters (state-based):** For filters requiring a `fit()` method, the hash is finalized before training, combining the class name, public attributes, and the unique hash of the training data.

Formally, for each filter instance f, the filter hash H_f is computed as:

$$H_f = \texttt{hash}(\texttt{class\_name}(f), \texttt{pub\_attributes}(f), H_{\text{train}})$$

where H_{train} is included only for trainable filters and corresponds to the hash of the training dataset used during `fit()`. For non-trainable filters, H_{train} is omitted.

Data provenance chain. LabChain maintains pipeline integrity through a composition of hashes. When a filter processes an input dataset with hash H_{in}, the resulting output dataset hash is:

$$H_{\text{out}} = \texttt{hash}(H_f, H_{\text{in}})$$

This mechanism forms a *chain of hashes* that encodes the complete provenance of the data as it flows through the pipeline. The initial dataset hash, defined by the user, serves as the root of this provenance chain.

Included and excluded factors. By design, cache keys depend only on: the filter class identity, its public configuration parameters, and upstream data hashes (which themselves depend on data content or user-provided identifiers). Keys do not depend on source code versions, library versions, runtime environments, or hardware characteristics, unless such factors are explicitly encoded as public parameters.

Cache control and invalidation. To manage caching behavior, LabChain provides a `Cached` wrapper filter offering fine-grained control through three parameters:

- `store_data`: Toggles persistence of processed output datasets.
- `store_model`: Enables or disables caching of fitted filter states.
- `overwrite`: Forces recomputation even when a matching hash exists in cache. When enabled, if `store_data` and/or `store_model` are also active, the newly computed results will replace the existing cached entries for that hash.

This mechanism supports iterative development workflows where researchers may need to force recomputation due to external changes not captured by configuration parameters (e.g., bug fixes in dependencies, hardware-specific numerical behavior). To mitigate the risk of stale results following internal logic updates or bug fixes, we recommend that users either explicitly set the `overwrite` flag to `True` for a single run or include a dedicated versioning parameter within the filter's public configuration to force a hash change.

Non-determinism and stochasticity. Filters relying on stochastic operations are reproducible only if sources of randomness (e.g., random seeds) are explicitly exposed as public configuration parameters. LabChain guarantees cache consistency based on the defined key structure but does not enforce bit-wise numerical reproducibility for filters that do not control their random state.

Validation. The correctness of cache hits and misses is validated through dedicated tests in the LabChain suite. Atomic behavior is verified in `tests/unit`, ensuring hashing determinism and cache hits under controlled changes to filter parameters. Collective behavior within complex experimental workflows is validated in `tests/integration`, ensuring state consistency and the reliable reuse of cached results when combining multiple framework elements, such as sequential and parallel pipelines.

2.3. Storage backends

LabChain's caching system is storage-backend agnostic through the `BaseStorage` abstraction, which defines a uniform interface for storing and retrieving cached data and models. The framework provides two built-in implementations:

- **Local filesystem:** Stores cached results in the local file system, suitable for individual researchers or single-machine workflows.
- **Cloud storage (S3):** Persists cached results to Amazon S3 or S3-compatible object storage, enabling collaborative environments where multiple teams share cached intermediate results across geographical locations.

Users can implement custom storage backends by extending `BaseStorage` and implementing its interface methods (`upload_file`, `download_file`, `check_if_exists`, etc.). This extensibility allows integration with institutional storage systems, alternative cloud providers, or specialized distributed filesystems without modifying LabChain's core logic. The storage backend is configured at the pipeline or filter level, allowing researchers to mix local and cloud storage within the same workflow. For example, expensive preprocessing steps can be cached to shared cloud storage for team-wide reuse, while intermediate debugging outputs remain local.

2.4. Dependency management: container and IoC

At the core of LabChain is the Container, a class-name-based manager implementing Inversion of Control (IoC) and Dependency Injection (DI). Components are registered via the `@Container.bind()` decorator. This decouples components, which are assembled by the container rather than depending on each other directly. This design is particularly useful when reconstructing pipelines with `build_from_dump()`, as the container can re-instantiate private components based on the serialized public configuration. While it introduces a learning curve, it reduces technical debt by exposing structural errors that would remain hidden in monolithic scripts.

2.5. Software design

Fig. 2 illustrates the UML design of LabChain's core system. The architecture emphasizes extensibility, modularity, and component reuse. Several classical design patterns are applied:

- **Factory/Abstract Factory**: `BaseFactory` enables dynamic registration and instantiation of plugins from textual identifiers, decoupling component creation from specific implementations.
- **Plugin/Strategy**: Filters, metrics, and storage systems inherit from `BasePlugin`, allowing dynamic substitution and experimentation with alternative processing strategies.
- **Composite**: `BasePipeline` can compose multiple filters and metrics into hierarchical pipelines, treating entire workflows as reusable entities.
- **Template Method**: Base classes (`BaseFactory`, `BasePlugin`, `BaseFilter`) define the general workflow (`fit`, `predict`, `evaluate`, `item_dump`), while subclasses implement specific steps.
- **Adapter/Wrapper**: `BaseStorage` provides unified access to diverse storage backends (local, S3), which can be extended with custom implementations.

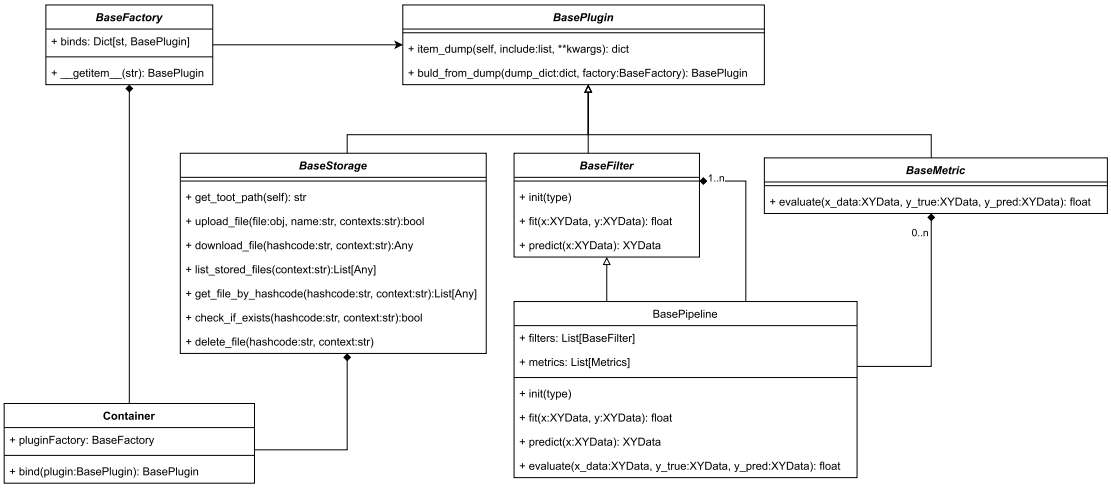


Fig. 2. General architecture of LabChain. Filters are organized into pipelines managed by a container. Each filter generates unique hashes that are stored in a content-addressable caching system.

This design allows independent testing, modular extensions, and portable reconstruction of pipelines using `build_from_dump`.

3. Illustrative examples

To illustrate the flexibility and power of the `LabChain` framework, we present a practical example focused on the early detection of signs of mental illness (based on temporal linguistic patterns). This example highlights three key capabilities: modular definition of filters, experiment serialization, and distributed execution of pipelines.

3.1. Pipeline construction and execution

As a use case, we implemented a pipeline inspired by research on temporal language analysis in social media [7]. The goal is to detect early signs of mental disorders based on users' posts (from Reddit). The data consist of sequences of messages, grouped by user, and organized into two classes (users diagnosed with depression vs control users).

The pipeline consists of two main phases:

- **Temporal feature extraction.** We use the Temporal Word Embeddings with a Compass model (`TWEC`), a variant of `Word2Vec`, to generate word embeddings that evolve over time [8]. By comparing these dynamic embeddings with their static versions, we obtain measures of individual semantic change for each word and user. The computation of measures of change is supported by different distance metrics (e.g., cosine, Euclidean, Chebyshev).
- **Classification and optimization.** The resulting representations score the evolution of the words in the vocabulary at different times. These features are fed into different classifiers – Support Vector Machines (`SVM`), Random Forest Classifiers (`RFC`), and K Nearest Neighbors (`KNN`)– which are automatically optimized using `GridSearch` with cross-validation. This allows prediction of signs of mental health based on the semantic evolution of language.

The output of the pipeline is a binary prediction of the user's mental health status, generated for each set of user's texts (the user's posts are organized in temporal *chunks*).

Custom filters. One of the advantages of `LabChain` is the ease of creating reusable filters. In this example, semantic distance matrices between embeddings (referred to as deltas) are precomputed and cached. These distances represent the movement of each word in each chunk relative to a canonical representation of that word (i.e., the degree to which the word shifts from its general usage). The following filter allows the selection at runtime of the distance metric to be applied:

```
@Container.bind()
class DeltaSelector(BaseFilter):
    def __init__(
        self,
        deltas_f: Literal[
            "cosine",
            "euclidean",
            "chebyshev",
            "jensen_shannon",
            "wasserstein",
            "manhattan",
            "minkowski",
        ] = "cosine",
    ):
        self.deltas_f = deltas_f
    def fit(self, x: XYData, y: XYData | None):
        pass
    def predict(self, x: XYData) -> XYData:
        stored_dok = x.value[self.deltas_f]
        if not isinstance(stored_dok, csr_matrix)
    :
            selected_matrix = stored_dok.tocsr()
        return XYData.mock(selected_matrix)
```

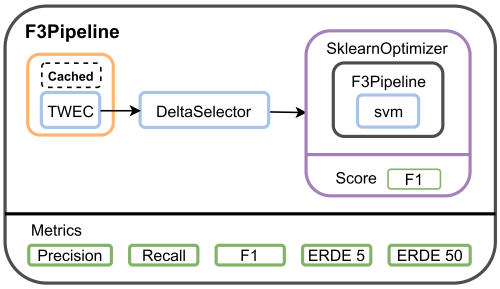


Fig. 3. Conceptual representation of the pipeline.

This filter acts as a view selector, allowing the user to easily switch the analysis metric without recalculating the underlying embeddings.

Pipeline design. Fig. 3 schematically represents the processing flow for this use case. All embeddings and deltas are generated in a single pass, cached, and then reused.

Advantages of the approach. This design demonstrates several key principles of the framework:

- **Efficiency.** The design's flexibility allows temporal embeddings and deltas to be generated in the initial filter, followed by a selector filter. This approach inherently reduces computation time. Furthermore, caching can significantly reduce the final computation time when using multiple classifiers, as would be the case when replicating the original paper's methodology.
- **Modularity.** Components are implemented as reusable and interchangeable filters. This architecture ensures easy adaptation to new datasets, evaluation metrics, or models without modifying the core structure.
- **Scalability.** The pipeline can be easily parallelized and executed in a distributed manner, regardless of data size or the number of experimental configurations.
- **Rich Evaluation.** The system supports both internal evaluation metrics (e.g., `F1` during optimization) and task-specific external metrics (e.g., `ERDE`, an error measure that penalizes late decisions about risk). Additionally, the caching mechanism enables detailed inspection of intermediate data, allowing for deeper and more informative evaluations at each pipeline stage.

Pipeline serialization. Pipelines can be easily exported to `JSON` using the `item_dump()` function, as shown below:

```
dumped_pipeline = pipeline.item_dump()
```

The output is a nested dictionary representing the pipeline structure and the public parameters of each component. This representation is easily serializable and shareable, enabling experiment storage, reproducibility, and distributed execution. Currently, only basic data types (e.g., str, int, list) are supported for filter parameters, excluding arbitrary objects.

3.2. Reproducibility and collaboration

The use of declarative pipeline representations enables seamless sharing and reconstruction of experiments. Pipelines serialize to JSON format, facilitating collaboration among team members and ensuring that computational workflows can be stored alongside datasets. This approach transforms pipeline definitions into executable documentation that captures not just the results, but the entire experimental process.

Reconstructing a pipeline is straightforward, a single function call restores the complete computational workflow. This capability enables researchers to systematically compare pipeline variants defined at different times, test alternative approaches, or apply identical processing to new datasets without risk of introducing subtle implementation differences.

```
from labchain.base import BasePipeline,
    BasePlugin
from typing import cast

reconstructed_pipeline: BasePipeline = cast(
    BasePipeline,
    BasePlugin.build_from_dump(
        dumped_pipeline, Container.pif
    )
)
reconstructed_pipeline.fit(X, y)
```

3.3. Distributed execution: exploring multiple pipelines in parallel

Classical distributed computing frameworks like Hadoop and Spark parallelize data processing by replicating the same computational logic across different data fragments. LabChain proposes a complementary strategy: parallelizing entire pipeline variants rather than data. This design proves particularly valuable in automated experimentation and systematic configuration exploration.

Consider a common research scenario: evaluating multiple classifiers (such as `SVM`, `RFC`, and `KNN`) on the same feature representations. With LabChain's distributed execution, different classifier configurations can be distributed across cluster nodes while sharing preprocessed data through the caching system. This eliminates redundant computation of expensive preprocessing steps and focuses computational resources on exploring classification and optimization strategies.

The following example demonstrates distributed pipeline execution. Each pipeline variant receives serialized configuration, executes independently on cluster workers, and the system automatically collects results into a unified output structure:

```
from labchain.pipeline import HPCPipeline
from labchain.plugins.filters import Knn,
    RandomForest, SVM

 f_experiment = lambda classifiers: F3Pipeline(
    filters=[
        TWEC(),
        DeltaSelector(),
        *classifiers
    ],
    metrics=[F1(), Precision(), Recall()]
)
pipeline = HPCPipeline(
    filters=[
        f_experiment([SVM()]),
        f_experiment([RandomForest()]),
        f_experiment([Knn()]),
    ],
    app_name="classifier_comparison",
    master="192.168.0.100:7077"
)
pipeline.fit(X, y)
predictions = pipeline.predict(X)
```

In this configuration, each branch of the `HPCPipeline` represents a complete experimental variant with a different classifier. The distributed system handles the complexity of scheduling, execution, and result aggregation, while cached intermediate results (embeddings and distance computations) are automatically shared across variants. Fig. 4 illustrates how these pipeline variants are distributed across cluster workers.

3.4. Hyperparameter optimization

Modern machine learning experiments require systematic exploration of hyperparameter spaces. This challenge has been central to automated machine learning research [9,10], where evaluating multiple pipeline configurations is considered essential for identifying robust and reproducible solutions. LabChain provides modular integration of multiple optimization strategies through a unified interface, supporting approaches ranging from exhaustive grid search to sophisticated Bayesian optimization, as well as visual tracking tools for monitoring experimental progress.

GridOptimizer. For problems with discrete hyperparameter spaces, LabChain includes a grid search optimizer that systematically evaluates all combinations of parameters defined through the `.grid()` method. This exhaustive approach works well when the search space is bounded and computational resources allow thorough exploration.[1]

OptunaOptimizer. When dealing with large or continuous hyperparameter spaces, Bayesian optimization offers a more efficient alternative. Through integration with Optuna [11], LabChain enables intelligent search strategies that learn from previous evaluations to guide exploration toward promising regions of the parameter space.[2]

WandbOptimizer. Systematic experimentation benefits from visual tracking and comparison of different configurations. Integration with Weights & Biases [12] provides real-time monitoring of optimization progress, comparison of multiple experimental runs, and support for remote collaboration among research team members.[3]

```
pipeline = F3Pipeline(
    filters=[
        StandardScaler(),
        Knn().grid({
            "n_neighbors": [2, 3, 4, 5, 6]
        })
    ],
    metrics=[F1(), Precision(), Recall()]
)
.splitter(
    KFoldSplitter(n_splits=5, shuffle=True)
)
.optimizer(GridOptimizer(scorer=F1()))

pipeline.fit(X, y)
```

[1] Tutorial: GridOptimizer example.
[2] Tutorial: OptunaOptimizer example.
[3] Tutorial: WandbOptimizer.

```
pipeline = F3Pipeline(
    filters=[
        PCA().grid({
            "n_components": [10, 100]
        }),
        SVM(kernel="rbf").grid({
            "C": [0.1, 10.0],
            "gamma": [0.001, 0.01]
        })
    ],
    metrics=[F1(), Precision(), Recall()]
)
.splitter(
    KFoldSplitter(n_splits=5, shuffle=True)
)
.optimizer(
    OptunaOptimizer(
        n_trials=50,
        direction="maximize",
        study_name="svm_optimization",
        load_if_exists=True,
        storage="sqlite:///optuna_studies.db"
    )
)

pipeline.fit(X, y)
```

```
pipeline = F3Pipeline(
    filters=[
        StandardScaler(),
        Knn().grid({
            "n_neighbors": [2, 3, 4, 5, 6]
        })
    ],
    metrics=[F1(), Precision(), Recall()]
)
.splitter(
    KFoldSplitter(n_splits=5, shuffle=True)
)
.optimizer(
    WandbOptimizer(
        project="mental_health_detection",
        sweep_id=None,
        scorer=F1()
    )
)

pipeline.fit(X, y)
```

SklearnOptimizer. For researchers already working within the scikit-learn [1] ecosystem, LabChain provides an optimizer that maintains compatibility with scikit-learn's validation infrastructure. This enables seamless integration with existing workflows while benefiting from LabChain's caching and pipeline management capabilities. The case study presented in this article demonstrates this optimizer in practice.

4. Impact

LabChain transforms how research teams develop, execute, and standardize computational experiments. Beyond technical improvements, it addresses fundamental challenges in scientific software: reproducibility, sustainability, and the psychological burden of experimental iteration.

Substantial improvement in experimental outcomes. To demonstrate practical utility, we re-implemented the experimental pipeline from prior work on early detection of mental health indicators [7]. The pipeline comprises three stages: (1) computing temporal representations using `TWEC` with multiple distance metrics, (2) selecting metrics via `DeltaSelector`, and (3) optimizing an `SVM` classifier. The initial stage alone requires one to several hours depending on dataset size.

Traditional implementations follow computational complexity N×(A+B+C), where N represents experimental variants and A, B, C represent pipeline stages. LabChain's caching reduces this to A+N×(B+C) by automatically reusing intermediate results. When exploring multiple classifiers on the same embeddings, this difference becomes transformative, expensive preprocessing executes once rather than repeatedly.

Beyond efficiency, caching provides psychological liberation. Research with dynamically typed languages like Python involves numerous subtle errors, type mismatches, configuration mistakes, and module incompatibilities, that often surface only after hours of computation. Automatic caching creates checkpoints that preserve progress, enabling researchers to explore alternative approaches without fear of losing work. This reduces both mental and computational costs of experimentation, encouraging the creative, iterative mindset essential to scientific discovery.

Preserved intermediate data also enables analytical depth. Researchers can study how different pipeline stages contribute to final performance without manually implementing inspection mechanisms. By embedding this capability into the framework, LabChain reduces technical debt while enhancing scientific focus.

The re-implementation using LabChain produced results demonstrably superior to those originally published, with some achieving state-of-the-art performance. This empirically confirms that well-designed tools not only improve development efficiency but can directly enhance scientific outcomes.

Environmental sustainability. Computational efficiency translates to environmental impact. Delta computation via `TWEC` on the depression dataset requires approximately three hours. The original study evaluated five models, without caching, this totals 15 hours per task. Extending to four tasks (depression, anorexia, self-harm, gambling) with multiple classifiers amounts to at least 48 hours of redundant computation.

Using conservative estimates (0.052 kg CO_2/hour), caching saves approximately 2.50 kg CO_2 across these experiments. Under higher emission assumptions (0.234 kg/hour), savings exceed 11 kg CO_2. These figures underscore both efficiency gains and contributions to sustainable research practices.

Empirical evidence. LabChain does not modify model architectures or algorithms, its impact lies in how experiments are structured. By reducing technical debt through modular organization, the framework creates mental space for deeper analytical thinking. When researchers are freed from managing ad-hoc scripts, tracking intermediate results manually, or reconstructing preprocessing pipelines, they can focus cognitive resources on identifying improvements, testing hypotheses, and refining experimental design.

This cognitive benefit manifested in tangible outcomes. During re-implementation of temporal language analysis for early risk prediction [7] using LabChain,[4] the modular structure immediately exposed a critical bug in the original corpus preparation. The preprocessing step failed to flatten the hierarchical data structure before feeding it to the word embedding model (TWEC). Instead of processing individual posts as separate documents, the entire corpus was concatenated into a single massive sequence. This structural error corrupted the training process: the model lost document boundaries, distorting word co-occurrence statistics and context windows essential for learning meaningful embeddings. The bug remained undetected across multiple experiments because monolithic preprocessing code obscured the data flow. LabChain's encapsulation of preprocessing as an explicit, testable filter made the structural corruption immediately visible during component validation.

Correcting this single structural error, combined with the systematic exploration enabled by LabChain's caching system, yielded the substantial improvements shown in Table 2. The results demonstrate that

[4] github.com/manucouto1/Temporal-Word-Embeddings-for-Early-Detection.

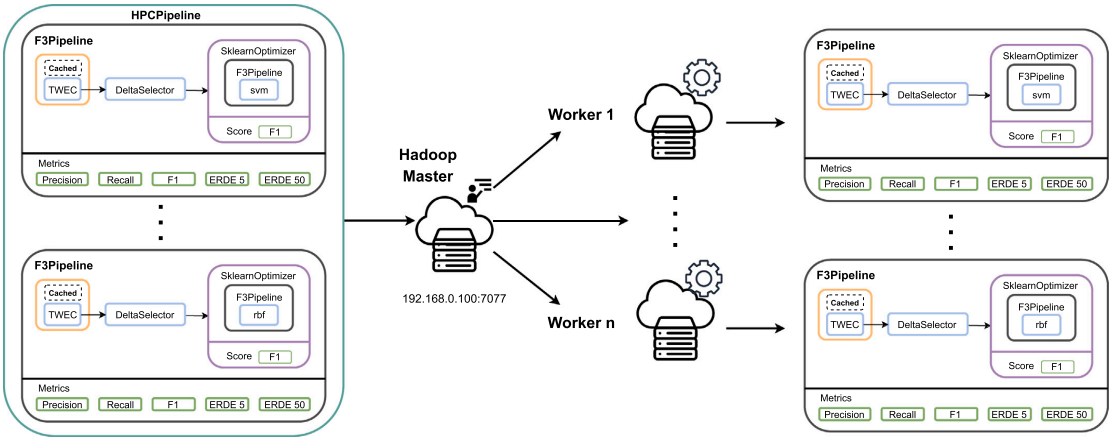


Fig. 4. Distribution of pipeline jobs in a Hadoop cluster. Each worker executes a complete pipeline variant, sharing cached preprocessing results.

Table 2
Performance comparison between original and LabChain implementations. Bold indicates best performance per dataset (lower ERDE values are better).

Dataset	Model	$ERDE_5$	$ERDE_{50}$	F1
Depression				
2017	LabChain	**12.44**	**7.66**	**0.50**
	Original	12.56	10.21	0.25
2018	LabChain	9.36	**5.56**	**0.46**
	Original	**9.09**	7.84	0.21
2022	LabChain	**6.57**	**3.81**	**0.46**
	Original	6.69	5.82	0.17
Anorexia				
2018	LabChain	**18.35**	**12.52**	**0.26**
	Original	21.40	15.23	0.22
2019	LabChain	**7.95**	**2.60**	**0.63**
	Original	8.17	6.30	0.27
Self-harm				
2020	LabChain	**19.96**	9.17	**0.62**
	Original	18.79	**16.37**	0.41
2021	LabChain	9.80	**5.42**	**0.33**
	Original	**9.02**	7.21	0.26
Gambling				
2022	LabChain	**3.35**	**0.88**	**0.38**
	Original	3.76	3.31	0.13
2023	LabChain	2.96	**0.41**	**0.88**
	Original	**2.92**	1.10	0.74

software engineering practices directly influence research outcomes: modularity makes errors visible that would otherwise remain buried in unstructured code.

Standardization and collaboration. Contemporary computational science lacks standardization in experimental development, impeding collaboration, traceability, and reproducibility. LabChain provides consistent infrastructure for defining, executing, and documenting experiments within a unified framework. This aligns with Baggen et al.'s [6] demonstration that code quality standards improve maintainability and development efficiency. By adopting modular design and separation of concerns, LabChain enhances collaboration among researchers with varying technical expertise.

The framework addresses growing demand for standardized methodologies in research software. Wainer et al. [13] emphasize that absent standards hamper tool interoperability, while Katz et al. [14] argue that community organizations transform research software culture toward sustainable, reproducible practices.

Adoption potential and future directions. LabChain's component-oriented design evokes established tools like LONI Pipeline [15], which accelerated reproducible neuroimaging research through visual interfaces and modular architecture. Similarly, it shares MLPro's [16] spirit of orchestrating complex workflows through consistent, extensible interfaces. The framework positions itself as a potential de facto standard for scientific experiments in AI, machine learning, and computational science.

Standardization through LabChain enables new research questions in scientific software engineering: how do coding styles, architectural quality, or code smells influence result reliability? Jerzyk et al. [17] show that development practices affect library security, while Kim et al. [18] demonstrate that poor design introduces vulnerabilities. Garlan et al. [19,20] argue that architectural patterns enable robustness evaluation. LabChain captures experimental architecture as versionable artifacts, facilitating such analyses.

Following Wright and Druta [21], LabChain serves as nexus between community and standardization. Effective standardization emerges from active communities sharing tools, conventions, and documentation. The framework enables groups to adopt common structures while maintaining flexibility, with reusable modules and versioned pipelines evolving into shared forms. This promotes participatory governance of computational knowledge, aligning with open science principles.

Current development focuses on graphical interfaces for users without programming expertise, similar to LONI's approach [15]. Future work includes federation of execution servers through actor-based architectures, enabling pipeline components to run across physical locations. Breivold et al. [22] advocate this architecture for large-scale service-oriented systems, LabChain could become the technological foundation for virtual laboratories and federated experimentation environments.

At a deeper technical level, future work will explore tighter integration with PyTorch, allowing neural network architectures to be expressed as cacheable filter pipelines. This would enable layer-wise pretraining with reuse of intermediate representations and selective parameter freezing for downstream tasks, extending LabChain's caching model to internal neural network components. In addition, ongoing work investigates mechanisms for coordinated execution over shared storage backends to mitigate race conditions under concurrent pipeline

usage. Finally, future releases are expected to incorporate versioned and persistent registries for plugins and serialized classes stored in shared backends, enabling remote pipeline execution without requiring full project source code at runtime. These directions remain subject to further validation across diverse execution environments before being considered for production use.

5. Conclusions

Scientific computing faces a software crisis that undermines reproducibility and wastes resources. LabChain addresses this through principled engineering: modular filters, hash-based caching, and JSON-serializable pipelines transform fragile research scripts into robust, reusable infrastructure. The framework demonstrates that software quality is not orthogonal to scientific quality, they are fundamentally linked.

The mental health detection case study validates this connection. Restructuring experiments with LabChain yielded both computational savings (12+ hours per task, reduced CO_2 emissions) and superior scientific outcomes, with performance improvements reaching 192.3% in some tasks. These gains emerged not from algorithmic innovation but from reduced technical debt: when researchers spend less cognitive energy managing code complexity, they invest more in hypothesis refinement and experimental insight.

This work challenges an implicit assumption in research software development: rapid prototyping requires sacrificing maintainability and exploration demands technical chaos. LabChain proves the opposite. Modularity, encapsulation, and automatic caching do not constrain creativity, they amplify it by removing friction from the experimental cycle. The framework's caching system acts as a psychological safety net, encouraging researchers to explore risky ideas without fear of losing progress.

Maturity and roadmap. LabChain is being actively maintained and employed daily in production research workflows at our research center (CiTIUS). Core components (filters, pipelines, Container, caching, data splitting, WandB optimizer) are production-ready. Experimental components (Sklearn/Optuna optimizers, distributed execution) are functional but have not been extensively tested. The API is stabilizing, with significant changes documented in GitHub releases. As an open-source project with limited resources, we strongly encourage community contributions.

The framework is publicly available with complete documentation at https://manucouto1.github.io/LabChain/. We invite the community not only to use LabChain, but also to embrace its underlying proposition: that applying software engineering principles to computational science is not a constraint on research, but an accelerant. The code we write shapes not only what we can compute, but what we can discover.

CRediT authorship contribution statement

Manuel Couto: Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Javier Parapar:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization. **David E. Losada:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the authors used Claude (Anthropic) to assist with improving language clarity, manuscript organization, and refining technical explanations. The tool was used to enhance readability and ensure consistent technical terminology throughout the document. After using this tool, the authors carefully reviewed, edited, and validated all content to ensure accuracy and alignment with the research objectives. The authors take full responsibility for the content of the published article.

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

MC and DEL thank the financial support provided by MICIU/AEI/10.13039/501100011033 (PID2022-137061OB-C22, supported by ERDF) and Xunta de Galicia-Consellería de Cultura, Educación, Formación Profesional e Universidades (ED431G 2023/04, ED431C 2022/19, supported by ERDF).

JP has received support from project PID2022-137061OB-C21 (MCIU/AEI/10.13039/5011000 11033, Ministerio de Ciencia e Innovación). He also thanks the financial support provided by the Consellería de Educación, Universidade e Formación Profesional, Spain (grant number ED481A-2024–079 and GRC ED431C 2025/49); and the European Regional Development Fund, which supports the CITIC Research Center.

Data availability

The LabChain framework is publicly available at https://github.com/manucouto1/LabChain. The reference implementation for this article is **v1.2.1**.

The mental health detection case study uses publicly available datasets from the eRisk shared tasks: Depression (2017, 2018, 2022), Anorexia (2018, 2019), Self-harm (2020, 2021), and Gambling (2022, 2023). These datasets can be requested from the eRisk organizers at https://erisk.irlab.org/.

The complete implementation of the case study, including all pipeline configurations and preprocessing code, is available at https://github.com/manucouto1/Temporal-Word-Embeddings-for-Early-Detection....

No new data were generated or analyzed in support of this research.

References

[1] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825–30.

[2] QuantumBlack AMC. Kedro: an open-source Python framework for creating reproducible, maintainable and modular data science code, 2019. [Accessed: 2025–05–23]; https://kedro.org.

[3] Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, Forster J, Lee S, Twardziok SO, Kanitz A, et al. Sustainable data analysis with snakemake. F1000Research 2021;10:33.

[4] Ali M. PyCaret: an open source, low-code machine learning library in Python, pyCaret version 1.0 Apr 2020. https://www.pycaret.org.

[5] Rieger M, Erdmann M, Fischer B, Fischer R. Design and execution of make-like, distributed analyses based on Spotify's pipelining package Luigi. arXiv preprint arXiv:1706.00955, 2017.

[6] Baggen R, Correia JP, Schill K, Visser J. Standardized code quality benchmarking for improving software maintainability. Softw Qual J 2012;20(2):287–307.

[7] Couto M, Perez A, Parapar J, Losada DE. Temporal word embeddings for early detection of psychological disorders on social media. J Healthc Inform Res 2025:1–30.

[8] Di Carlo V, Bianchi F, Palmonari M. Training temporal word embeddings with a compass. In: Proceedings of the AAAI conference on artificial intelligence, vol. 33. 2019. p. 6326–34.

[9] Feurer M, Klein A, Eggensperger K, Springenberg J, Blum M, Hutter F. Efficient and robust automated machine learning. Adv Neural Inf Process Syst 2015; 28.

[10] Olson RS, Bartley N, Urbanowicz RJ, Moore JH. Evaluation of a tree-based pipeline optimization tool for automating data science. In: Proceedings of the genetic and evolutionary computation conference 2016; 2016. p. 485–92.

[11] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: a next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery &. data mining; 2019. p. 2623–31. https://doi.org/10.1145/3292500.3330701

[12] Biewald L. Experiment tracking with weights and biases, software available from wandb.com; 2020. https://www.wandb.com/.

[13] Wainer GA, Al-Zoubi K, Hill DR, Mittal S, Martín JLR, Sarjoughian H, Touraille L, Traoré MK, Zeigler BP. An introduction to DEVS standardization. In: Discrete-Event Modeling and Simulation, CRC Press; 2018. pp. 393–425.

[14] Katz DS, McInnes LC, Bernholdt DE, Mayes AC, Hong NPC, Duckles J, Gesing S, Heroux MA, Hettrick S, Jimenez RC, et al. Community organizations: changing the culture in which research software is developed and sustained. Comput Sci Eng 2018;21(2):8–24.
[15] Rex DE, Ma JQ, Toga AW. The LONI pipeline processing environment. Neuroimage 2003;19(3):1033–48.
[16] Arend D, Diprasetya MR, Yuwono S, Schwung A. MLPro: an integrative middleware framework for standardized machine learning tasks in Python. Softw Impacts 2022;14:100421.
[17] Jerzyk M, Madeyski L. Code smells: a comprehensive online catalog and taxonomy. In: Developments in information and knowledge management systems for business applications, vol. 7: Springer; 2023. pp. 543–76.
[18] Kim S, Lee H. Software systems at risk: an empirical study of cloned vulnerabilities in practice. Comput & Secur 2018;77:720–36.
[19] Garlan D, Shaw M. An introduction to software architecture. In: Advances in software engineering and knowledge engineering. World Scientific; 1993. pp. 1–39.
[20] Garlan D. Software architecture: a travelogue. In: Future of software engineering proceedings; 2014. p. 29–39.
[21] Wright SA, Druta D. Open source and standards: the role of open source in the dialogue between research and standardization. In: 2014 IEEE globecom workshops (GC Wkshps). IEEE; 2014. p. 650–5.
[22] Breivold HP, Crnkovic I, Larsson M. A systematic review of software architecture evolution research. Inf Softw Technol 2012;54(1):16–40.

CHAPTER 5

RESULTS: ARTICLES UNDER REVIEW

This chapter presents a single contribution that is currently under review for publication in Knowledge and Information Systems (KIS), Springer [19]. Section 5.1 presents a comprehensive study of word embedding models for measuring topic coherence, which addresses the evaluation challenges identified in the preceding publications.

5.1 Publication III: A Study of Word Embedding Models for Measuring Topic Coherence

Authors: Manuel Couto, Javier Parapar, David E. Losada
Submitted to: Knowledge and Information Systems (KIS) [19]
Status: Under review
Keywords: Topic Coherence, Embeddings, Metrics

A Study of Word Embedding Models for Measuring Topic Coherence

Manuel Couto[1], Javier Parapar[2], David E. Losada[1]

[1]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain.

[2]Information Retrieval Lab, CITIC, Universidade de A Coruña, A Coruña, Spain.

Contributing authors: manuel.couto.pintos@usc.es; javier.parapar@udc.es; david.losada@usc.es;

Abstract

Topic modeling has emerged as a crucial tool in the field of natural language processing, enabling the automatic discovery of latent structures in large textual corpora. However, determining the quality of the topics remains a significant challenge, particularly in measuring the coherence of the top words of the extracted topics. Early efforts relied on human judgments, but these approaches are resource-intensive. Automated coherence metrics have since been developed. For example, some measures exploit word co-occurrence, while other methods are grounded in distributional semantics (e.g., employing word embeddings). In this study, we thoroughly explore the application of embedded representations to evaluate the quality of topics. While a number of isolated studies have analyzed the role of specific word representation techniques for measuring topic coherence, a complete picture of their effectiveness is still lacking. This work brings together different embedding-based approaches, including Word2Vec, FastText, GloVe, and BERT, which had been studied separately, and extends prior research by incorporating additional models such as RoBERTa, ALBERT and MPNET. Topic coherence is measured by computing similarity scores between word embeddings, thus obtaining rich semantic associations that traditional measures may overlook. Our analysis demonstrates that these methods are as effective as, and often surpass, classical coherence measures. Our results contribute to a growing body of research advocating for advanced semantic representations as robust alternatives to traditional approaches in evaluating topic model coherence.

Keywords: Topic, Coherence, Embeddings, Metrics

1 Introduction

Topic modeling [1–4] is an unsupervised learning technique employed to discover latent topics within a collection of documents. Each topic is typically represented as a probability distribution over words, reflecting the relevance of each word within the topic. The most representative words of a topic –those with the highest probability, commonly referred to as *top words*– facilitate the presentation of topics to users interested in exploring the document corpus and analyzing the resulting themes. Evaluating the quality of topics generated by automated topic modeling systems is challenging due to the subjective and complex nature of what constitutes a "good" topic. The top words of a good topic are generally assumed to be interpretable, coherent, and meaningful. For example, a coherent topic in the domain of computing might include words such as `"AI", "computer", "network", "parallelism", "algorithm"`, which are semantically related and form a clear, interpretable theme. In contrast, an incoherent topic might contain top words like `"AI", "novel", "tomato", "democracy", "fast"`, where the lack of a clear semantic connection makes interpretation difficult.

Assessing the topic's top words based on human judgments is costly and lacks scalability. This prompted the development of automated coherence metrics. Among these metrics, those based on word co-occurrence have been widely used; however, they often fail to capture the semantic richness of word relationships in complex contexts [5–7]. Despite these criticisms, many recent studies proposing novel topic modeling approaches -—including neural topic models, dynamic generative models, and multilingual extensions—- still rely on classical co-occurrence metrics such as NPMI and UCI to evaluate coherence [4, 8–11]. Word embedding representations generated by classic models like Word2vec, GloVe, and FastText or by more recent contextual approaches like BERT, RoBERTa, ALBERT or MPNET offer a promising alternative, as these types of word representations more accurately model semantic relationships between words. Previous work [7, 12–15] has explored metrics based on this model's embeddings for assessing topic coherence, leveraging representations such as FastText, word2vec, GloVe, and contextual models like BERT. However, a complete picture of their effectiveness is still lacking. Moreover, the potential of input embeddings or lower-layer representations from contextual models for topic coherence evaluation remains unexplored.

Our study aims to fill this gap by comparing traditional coherence measures and newer embedding-based metrics. We formally evaluate multiple coherence methods by computing the correlation between the automated estimates and ground truth assessments made by humans. Several embedding alternatives incorporated into our study have never been empirically tested for measuring topic coherence. Our main goal is to provide an in-depth, comparative evaluation of coherence metrics to guide future research and development. Our survey demonstrates that coherence estimation using embeddings can be as effective as, or even better than, traditional metrics based on co-occurrence estimates.

We formulate the following research questions to guide our study:

- **RQ1:** What is the relative performance of different classical coherence metrics?

- **RQ2:** How does the corpus used to estimate word-to-word co-occurrence affect classical metrics?
- **RQ3:** How do novel word embedding-based coherence metrics compare?
- **RQ4:** Are new word embedding-based metrics competitive against classical approaches in topic coherence evaluation?

2 Related Work

The evaluation of topic quality has been a recurring area of interest, commonly addressed through analyzing key top words and their coherence. Early approaches relied on human annotators, but this method is resource-intensive and costly [16–18].

Since 2009, automated metrics have begun to emerge to assess topic quality. AlSumait [19] proposed an automated metric to identify and classify irrelevant or junk topics. Their method leveraged aspects such as the uniformity of word distributions within a topic (W-Uniform), the detection of vacuous or non-informative word distributions (W-Vacuous), and the analysis of background distributions (D-BGround) to determine topic relevance. These estimates, grounded in statistical and distributional criteria, enabled automatic detection of low-quality topics.

A year later, Newman and colleagues introduced UCI [20], a metric designed to assess topic quality based on semantic coherence. UCI analyzes the co-occurrences between each pair of top words in a topic. Co-occurrence is estimated over a reference corpus using a 10-word sliding window. Specifically, UCI calculates the Pointwise Mutual Information (PMI) for each pair of top words, assuming that words frequently co-occurring in actual documents are more likely to form human-interpretable coherent topics. Building on this method, Mimno et al. [21] proposed the UMass metric, which focuses on the conditional probability of co-occurrence between a topic's top words. This approach calculates the document frequency of each word and the co-document frequency of word pairs within the same document, eliminating the reliance on an external reference corpus. The resulting pairwise log conditional probability (LCP) favors topics whose most probable terms frequently co-occur. The metric's quality was evaluated with ROC curves that compare the automated topic estimates with human-generated assessments.

Aletras and Stevenson [22] applied distributional semantics to develop a normalized variant of PMI, known as NPMI. Their approach calculates the average distance between vector representations of word pairs within a topic, using similarity metrics such as Cosine, Dice, and Jaccard. Topic coherence estimation was based on word pair co-occurrence, and NPMI was a pioneer in what would later be known as indirect confirmation measures. This innovation provided a more precise and effective method for evaluating topic quality. Continuing to advance the field, Röder et al. [23] defined a framework that synthesizes and organizes the characteristics of existing coherence metrics. This study also introduced two new coherence metrics, CV and CP, and evaluated them through correlation measures between automated quality estimates and human-generated rankings.

During this period, significant advancements in word embeddings paved the way for new topic coherence evaluation techniques based on vectorial representations. In 2013,

Word2Vec [24] revolutionized word vector representations by introducing distributed embeddings learned from neural network architectures. Trained with a word or context prediction objective (Skip-gram or Continuous Bag of Words), Word2Vec captured semantic relationships with unprecedented accuracy. GloVe [25] emerged shortly after as a powerful alternative, leveraging global co-occurrence matrix factorization to produce dense word representations. Later, FastText [26] introduced subword embeddings, enhancing the expressiveness of dense representations and further expanding the applicability of word vectors.

As word embeddings gained traction in natural language processing, several studies explored their potential for measuring topic coherence. These studies followed alternative methods for leveraging word vector representations to improve coherence estimation. In 2016 Nikolenko [12] expanded upon embedding-based approaches by introducing Distributed Word Representations (DWR). This method is founded on the idea that a coherent topic should be well-clustered in the semantic space, meaning that a topic's top words should be close together when mapped into an embedding space. By utilizing embeddings from models like Word2Vec or GloVe, pre-trained on reference corpora, the method calculates the mean distance between top words using cosine, L_1, and L_2 distances, providing a direct, geometry-based evaluation of coherence. Fang et al. [13] took a different approach when exploiting Word2Vec and GloVe to evaluate topics extracted from Twitter and Wikipedia. Unlike Nikolenko's method, which focused on spatial clustering, Fang et al. compared embedding-based metrics with traditional approaches like PMI and LSA through two complementary evaluations: (1) assessing how well automated metrics ranked topics compared to human reviewers, and (2) measuring the agreement between automated metrics and human judgments when evaluating topic pairs. Their study highlighted the strengths and weaknesses of embedding-based techniques across different domains. In 2017, Ramrakhiyani et al. [15] introduced a bucket-based approach to topic coherence evaluation using word embeddings, such as GloVe. By applying Singular Value Decomposition (SVD) for dimensionality reduction and integer linear programming (ILP) to optimally group a topic's top words into "buckets", these authors posited that larger, more cohesive buckets, indicate higher semantic coherence. This method outperformed earlier metrics on datasets like NYT, 20NG, and Genomics. In 2019, Belford and Greene [14] conducted a comparative study of embedding techniques for topic coherence. Using datasets from The Guardian, this team evaluated Word2Vec and FastText for measuring mean distances between top words, focusing on selecting the optimal number of topics (k). More recently, Doogan and Buntine [27] revisited the reliability of traditional and distributional coherence metrics, questioning their practical effectiveness in capturing semantic interpretability.

With the advent of transformer architectures, groundbreaking models like BERT [28], RoBERTa [29], and ALBERT [30] have reshaped NLP by introducing context-aware, dynamic embeddings. These models differ from earlier approaches, such as Word2Vec and GloVe, by capturing deeper contextual nuances. Despite their success in various NLP tasks, their application for topic coherence evaluation remains underexplored.

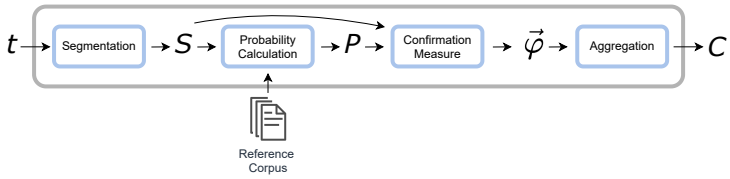


Fig. 1 Diagram illustrating the steps to compute the coherence (C) of a topic (t) [31]

While notable progress has been made in developing metrics and methodologies for topic coherence evaluation, prior studies have primarily made isolated proposals that were tested under specific experimental conditions. Our contribution lies in systematically comparing multiple word embedding models, shedding light on their relative performance and contextualizing them against classical approaches.

3 Classical Coherence Metrics

In the literature, numerous metrics have been proposed to evaluate the interpretability and distinctiveness of topics. Among them, classical coherence metrics are measures grounded in co-occurrence statistics while more recent methods are based on embeddings and other neural-induced representations. For the description of coherence metrics, we mainly rely on the framework introduced in [20, 31], which offers a comprehensive schematic representation and a common language for explaining and constructing classical coherence metrics. Under this framework, coherence metrics share a common structure consisting of four stages: **Segmentation, Probability Estimation, Confirmation Measures, and Aggregation** (see Figure 1).

The process begins with the set W of top words defining a given topic t. This is an ordered set, denoted as $W = \{w_1, w_2, ...\}$, where words are ranked in descending probability according to the topic's distribution. Different variations exist for each stage to address specific objectives.

During the **segmentation phase**, the topic's set of top words is divided into smaller subsets or segments. The segmentation process generates pairs of subsets, where the first is denoted as W' and the second as W^*. These pairs are later used in different comparison steps. The most intuitive segmentation method, S^{one}_{one}, forms all unique word pairs from the topic. For example, for a topic represented by a set of words $\{Dog, Cat, Snake\}$, this method would generate the pairs $(Snake, Dog)$, (Cat, Dog) and $(Cat, Snake)$. This strategy is adopted by the UCI and NPMI metrics, which compare all word pairs in W individually.

The S^{one}_{all} segmentation method compares each word with all other words in the set (subject to the constraint $W' \cap W^* = \emptyset$). The CV metric follows the segmentation method S^{one}_{set}, where each word is compared with all words in W. Unlike S^{one}_{all}, this method does not require W' and W^* to be disjoint, allowing a word to appear in both subsets.

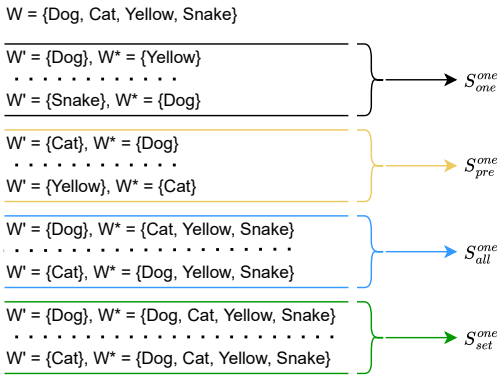


Fig. 2 Examples of segmentation techniques applied to a topic whose top words are "Dog", "Cat", "Yellow", and "Snake".

In the case of UMass, its most common variant employs a segmentation method denoted as S^{one}_{pre}, where words are ordered by their probability within the topic. Comparisons are made only between a word and its preceding words. Figure 2 shows an example to clarify these segmentation methods further. Formally, the segmentation alternatives are defined as follows:

$$S^{one}_{all}(W) = \Big\{(W', W^*)|W' = \{w_i\}; w_i \in W; W^* = W \setminus \{w_i\}\Big\} \quad (1)$$

$$S^{one}_{set}(W) = \Big\{(W', W^*)|W' = \{w_i\}; w_i \in W; W^* = W\Big\} \quad (2)$$

$$S^{one}_{one}(W) = \Big\{(W', W^*)|W' = \{w_i\}; W^* = \{w_j\}; w_i, w_j \in W; i \neq j\Big\} \quad (3)$$

$$S^{one}_{pre}(W) = \Big\{(W', W^*)|W' = \{w_i\}; W^* = \{w_j\}; w_i, w_j \in W; i > j\Big\} \quad (4)$$

The **probability estimation phase** defines how probabilities are derived from the reference corpus. These probabilities are a basis for evaluating the coherence between segments (W', W^*). UMass utilizes a Boolean variant, known as P_{bd}, which estimates probabilities based on the co-occurrence of words within documents. It ignores word repetitions within a document and the distance between word occurrences. CV, UCI, and NPMI follow another probability estimation strategy, P_{sw} (Boolean sliding window), where word co-occurrences are evaluated within fixed-size sliding windows across the textual corpus. For instance, $P_{sw(110)}$, used in CV, operates with windows of 110 tokens, shifting the window by one token per step to create virtual documents for co-occurrence estimation. The probability of each word is then computed as the

ratio of virtual documents in which it appears divided by the total number of virtual documents.

Once the probabilities are obtained, the **confirmation phase** estimates how strongly W^* supports W', where higher support implies higher coherence. Direct confirmation measures constitute the core of coherence metrics because they define how the relationship between words is modeled. This modeling is essential to capture the nuanced statistical dependencies that underlie the semantic association. Specifically, UCI computes a log-ratio measure (m_{lr}) to emphasize relative changes in co-occurrence likelihood; NPMI employs a normalized log-ratio measure (m_{nlr}) to standardize values across varying frequencies (allowing comparison between different word pairs) and UMass calculates a log-conditional probability measure (m_{lc}) that focuses on the predictive strength of one word for the occurrence of another.

Indirect confirmation metrics extend direct confirmation approaches, enabling coherence estimation between words that do not frequently co-occur but appear in similar contexts. For instance, "happy" and "sad" may not frequently co-occur but often appear alongside the word "feel". This suggests that in a set W, direct confirmations between words in W' and those in W^* exhibit similarities across all pairwise confirmations [31].

Finally, during the **aggregation phase**, all confirmation values $\vec{\varphi} = \{\varphi_1, \cdots, \varphi_{|S|}\}$ computed from the available segment pairs $S_i \in S$ are consolidated into a single coherence score. For instance, metrics such as CV, UMass, UCI, and NPMI typically employ the arithmetic mean (σ_a) for aggregation. Alternative approaches have also been explored. For example, Newman et al. [20] demonstrated that the median can reduce the influence of outliers. Additionally, methods like the *word buckets* approach aggregate confirmation values by grouping words into buckets based on criteria such as frequency or semantic similarity, and then computing bucket-level statistics [15].

The upper block of Table 1 compiles and organizes the main characteristics of four classical coherence metrics (CV, NPMI, UMASS and UCI) according to this four-stage framework (segmentation, probability, confirmation, and aggregation).

4 Embedding-Based Metrics

Our study also explores the effectiveness of coherence metrics based on embeddings from advanced language models. These metrics leverage pre-trained language models that capture semantic relationships between words within a vector space. Specifically, we use the following representations: GloVe, Word2Vec, FastText, BERT, RoBERTa, ALBERT and MPNET [24–26, 32–35]. These models were selected to cover two distinct families of word representations. On one hand, we consider three static non-contextualized embeddings –Word2Vec, GloVe, and FastText– that differ in their underlying technology and training methodology: Word2Vec is widely recognized as a seminal word representation model, GloVe captures global co-occurrence information, and FastText leverages subword information. On the other hand, we consider several transformer-based models –BERT, RoBERTa, ALBERT, and MPNET– which represent advanced contextualized embeddings. BERT serves as a referential contextual representation, while RoBERTa and ALBERT introduced modifications in BERT's

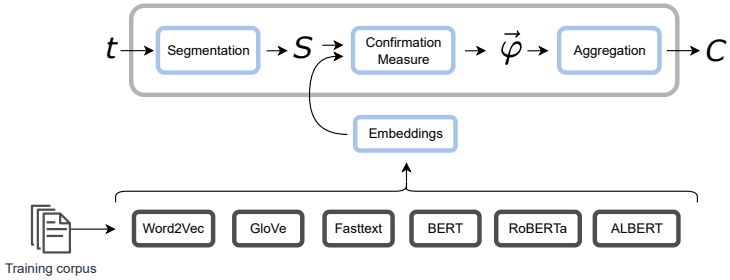


Fig. 3 Diagram illustrating the computation of coherence (C) for a topic (t) based on embeddings

training strategy (with ALBERT providing an optimized, parameter-reduced version of BERT). MPNET is a more recent model that combines strategies from masked language modeling and permutation language modeling. By including models from both families, our study not only extends previous empirical studies but also integrates new approaches that have shown improved performance across various evaluation metrics.

These embeddings can fit within a topic coherence framework similar to the one used to describe classical topic coherence metrics. Specifically, we can view the embedding representations as substitutes for the probability calculation stage. Rather than estimating co-occurrence probabilities, we measure the similarity between dense word representations from embedding models. Figure 3 illustrates how embedding models integrate into the coherence estimation process.

4.1 Embedding Models

In this section we briefly introduce the embedding models used in our study. The selection of these seven representative models constitutes a comprehensive evaluation setup, allowing for a thorough comparison across different families of representations.

Word2Vec [24] was a pioneering approach in effectively and efficiently generating dense word representations, significantly improving results in semantic and syntactic similarity tasks compared to earlier neural network-based models [36–39]. Word2Vec simplified the architecture of previous neural network language models (NNLMs) [36, 37] by removing the hidden layer, eliminating its non-linearity and significantly improving computational efficiency. This improvement allowed the model to be trained on much larger datasets more efficiently. Two alternative architectures were proposed: the Continuous Bag-of-Words model (CBOW) and Continuous Skip-gram. In both cases, as a by-product of the training process on large text collections, dense word representations (embeddings) are obtained. In the resulting vectorial space, semantically similar words are placed close to each other. We experiment here with the word2vec-google-news-300 model.

GloVe, introduced in 2014 [25], aimed at combining the strengths of two major language model families proposed up to that point. On one hand, it leveraged

matrix factorization approaches that generate low-dimensional word representations. On the other hand, it incorporated local (window-based) models that predict words within constrained contexts. To achieve this, GloVe factorizes a global word-word co-occurrence matrix, which is built using statistics gathered from local context windows. This allows the model to capture both global corpus-wide statistical relationships and local contextual information. We experiment here with the glove-wiki-gigaword-300 model.

FastText [26] aimed to improve dense word representations by modeling subwords, which is convenient for dealing with complex corpora, extensive vocabularies, rare words, jargon, or misspellings. FastText represents each original word with sets of subwords. We experiment here with the fasttext-wiki-news-subwords-300 model.

BERT [28] represented a significant advancement in the realm of language technologies. One of the main problems of previous models was their limited ability to disambiguate polysemic words and to capture how word meaning varies depending on the context. For example, "apple" does not have the same meaning when referring to a fruit as it does in a technology-related context. However, models like Word2Vec ignored context and generated the same representation for each word, regardless of its surrounding context.

BERT introduced the ability to generate dense token-level representations that depend on the context in which the token appears. This means that a token in one context will have a different representation than the same token in another context, allowing the model to capture semantic nuances dependent on word usage. This innovation was achieved through the attention mechanism and the bidirectional consideration of context (handling previous and subsequent words) [28].

Although prior research had already explored attention mechanisms combined with recurrent neural networks (RNNs), BERT removed the dependency on these architectures. Instead, it employed an algebraic operation called scaled dot-product attention, which can be efficiently computed using matrix operations. This feature simplified the architecture and enabled the exploitation of the inherent parallelism in specialized hardware, such as GPUs. This led to accelerated computational times for training and inference, thus establishing the foundation for highly scalable and effective models, such as BERT itself. We experiment here with the bert-base-uncased model.

RoBERTa (Robustly Optimized BERT Pretraining Approach) [33] was introduced as an improvement over BERT, focusing on optimizing its pretraining process to maximize performance in downstream tasks. Although it retains BERT's base architecture, RoBERTa implements several key modifications: it removes the Next Sentence Prediction (NSP) task, significantly increases training data and training time, and utilizes larger batch sizes. Additionally, it introduces advanced strategies for dynamic masking sampling, improving effectiveness in masked token prediction. These optimizations contribute to improved performance on downstream tasks without requiring changes to the original architecture. We experiment here with the roberta-base model.

ALBERT (A Lite BERT) [34] is a variant of BERT designed to reduce computational and memory requirements without significantly compromising performance. The main difference between ALBERT and BERT lies in its optimized architecture. ALBERT introduces two key techniques: embedding matrix factorization and

parameter sharing. Factorization reduces the size of the embedding matrices by decomposing them into two smaller matrices, thus decreasing the number of parameters needed to represent the vocabulary. On the other hand, parameter sharing reuses the same parameters across all model layers, drastically reducing the total number of parameters. These innovations allow ALBERT to be significantly lighter and more efficient than BERT, making it more suitable for resource-constrained environments while maintaining competitive performance in natural language processing tasks. We experiment here with the albert-base-v2 model.

MPNet [35] is a pre-training model for language understanding that leverages the advantages of masked language modeling and permuted language modeling while addressing their shortcomings. BERT independently predicts masked tokens using bidirectional context, while XLNet captures token dependencies via permutation but omits full positional context [40]. MPNet, instead, conditions its predictions on both the complete sequence's positional information and the dependencies among masked tokens. This is achieved through a unified training objective incorporating two-stream self-attention and position compensation, thereby reducing the discrepancy between pre-training and fine-tuning. Consequently, MPNet leverages these mechanisms to integrate information from both modeling paradigms, aiming to learn more comprehensive contextual representations. This model achieved state-of-the-art results on benchmarks like GLUE at the time of its publication while maintaining computational efficiency [35]. We experiment here with the mpnet-base model.

4.2 Segmentation with Embeddings

For the coherence calculation with embedding models, we first segment the top words of each topic following a S^{one}_{one} strategy. Then, vector representations (embeddings) generated by the pre-trained models are used, effectively replacing the role of probability estimations in classical metrics. For the confirmation process, we use vector similarity calculations (detailed below). This will allow us to exploit the capability of dense representations to measure how words within a topic "reinforce" each other.

4.3 Confirmation with Embeddings

For each embedding model, we measure the direct confirmation between the top words of the topics using their vector representations. In this stage, we evaluate how the words within a topic confirm each other by calculating the similarity between the vector representations of these words using the following metrics:

- **Cosine Similarity** (m_{cos}): Measures the angular similarity between vectors in the vector space. It is widely used in information retrieval and semantic search tasks.

$$m_{cos} = \frac{\vec{x} \cdot \vec{y}}{||\vec{x}|| \cdot ||\vec{y}||} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}} \tag{5}$$

- **Dot Product** (m_{dot}): Measures the similarity between two vectors by calculating their dot product. Unlike cosine similarity, it is not normalized by vector magnitudes, making it sensitive to the length of the embedding vectors.

$$m_{dot} =< \vec{x}, \vec{y} >= \sum_{i=1}^{n} x_i \cdot y_i \tag{6}$$

where $\vec{x} =< x_1, ..., x_n >$ and $\vec{y} =< y_1, ..., y_n >$ are embedding vectors representing two words.

The choice of these two variants is intentional and aims to assess coherence metrics under complementary similarity regimes. Cosine similarity normalizes vector magnitudes and emphasizes angular proximity, while dot product preserves vector norms and therefore reflects both direction and magnitude. Different embedding models induce distinct norm distributions (as a consequence of their training objectives and corpora) and, thus, comparing normalized and non-normalized similarities allows us to analyze how vector normalization affects coherence estimation across heterogeneous representations.

4.4 Optimal Word Buckets

A specific case of embedding-based metrics is TBuckets [15]. This approach is inspired by how humans analyze topic coherence. When a reviewer examines the top words of a topic, she starts with the first word and an empty "bucket". As she reviews the rest of the words, she must decide whether each new word fits into an existing bucket or a new bucket should be created. In the end, the words of the topic are distributed into different buckets. A distribution with one dominant bucket and a few smaller ones is considered indicative of high coherence, while fragmentation into multiple buckets of similar size suggests lower coherence.

To perform this grouping, TBuckets extracts the principal eigenvectors from the embedding matrix of the top words. These eigenvectors capture the underlying thematic axes, representing the main topic of the word set. A topic is considered coherent if most of the topic's words align with the principal eigenvector while other eigenvectors reflect minor thematic variations.

TBuckets uses Singular Value Decomposition (SVD) to decompose the semantic structure of the set and Integer Linear Programming (ILP) to optimize the assignment of words to buckets. During the optimization process, the cosine similarity between the word embeddings and the eigenvectors is computed to guide the assignment, dynamically adjusting the word distribution until achieving the best alignment, subject to a set of restrictions.

5 Summary of Coherence Metrics

After reviewing classical and embedding-based metrics, we now present a table summarizing different metrics and how they were evaluated in the literature (see Table 1). The table includes the characteristics of each model regarding segmentation, probability calculation or embedding representation, confirmation, and aggregation. Additionally, we report the data used for experimentation, which came from various sources like X or Wikipedia, and how each measure was evaluated. For example, some proposals

Metric	Segm.	Prob	Conf	Agg	Corpus	Eval
CV [23]	S^{one}_{set}	$P_{sw(110)}$	$\tilde{m}_{cos(nlr,1)}$	σ_a	Wikipedia & Same Corpus	Spearman
NPMI [22]	S^{one}_{one}	$P_{sw(10)}$	m_{nlr}	σ_a	Wikipedia	Spearman
UCI [20]	S^{one}_{one}	$P_{sw(10)}$	m_{lr}	σ_a	Wikipedia	Spearman
UMass [21]	S^{one}_{pre}	P_{bd}	m_{lc}	σ_a	Same Corpus	Kendall
Metric	Segm.	Emb	Conf	Agg	Corpus	Eval
TBuckets [15]	S^{one}_{one}	GloVe	$d_{cos(eign)}$	$\sigma_{buckets}$	Wikipedia	Pearson
DWR_{w2v} [12]	S^{one}_{one}	Word2Vec	$d_{cos,L_1,L_2,\xi}$	σ_a	Russian Web Corpus	AUC
G_W_WE [13]	S^{one}_{one}	GloVe	d_{cos}	σ_a	Wikipedia	Ranks & Agreement
G_T_WE [13]	S^{one}_{one}	GloVe	d_{cos}	σ_a	Twitter (X)	Ranks & Agreement
T_WE [13]	S^{one}_{one}	Word2Vec	d_{cos}	σ_a	Twitter (X)	Ranks & Agreement
Emb_{Fstxt} [14]	S^{one}_{one}	FastText	d_{cos}	σ_a	The Guardian & Wikipedia	k-topics
Emb_{w2v} [14]	S^{one}_{one}	Word2Vec	d_{cos}	σ_a	The Guardian & Wikipedia	k-topics

Table 1 Classical coherence metrics (top block) and embedding-based metrics (bottom block). For each strategy, the type of Segmentation used (Segm.), the Probability (Prob.) or Embedding (Emb.) calculation, the Confirmation measure (Conf), Aggregation strategy (Agg), the corpus used in the experimentation, and the evaluation strategy (Eval) are shown.

were evaluated using Kendall correlation between reference topic rankings and rankings produced by the metrics, while other proposals used Spearman correlation. In other cases, AUC measures or other strategies were employed. Full details of these studies are available in the article referenced in the first column of each row.

This table shows that many alternatives (e.g., RoBERTa, ALBERT or MPNET embeddings) were never evaluated for measuring topic coherence.

6 Experimental Setup

In this section, we detail the procedure to evaluate the topic coherence metrics, comparing classical coherence metrics with new metrics based on embeddings. As shown in the previous table, these metrics have been independently evaluated using different corpora and evaluation mechanisms. However, there is no systematic comparison in the literature that evaluates all variants –including multiple embedding models– under homogeneous conditions. We considered the following **topic coherence metrics**:

- **Classical Metrics:** CV, UMass, UCI, and NPMI, which are based on word co-occurrence relationships and require a reference corpus.
- **Embedding-based Metrics:** Here, we consider representations generated by GloVe, Word2Vec, FastText, BERT, RoBERTa, ALBERT and MPNET. The top words in the topics are individual lexical items with no associated sentential or discourse-level context (i.e., no surrounding words). Consequently, for transformer-based models, we restrict our analysis to input (token) embeddings, treating each top word as an independent unit. This design choice is deliberate and aims to preserve methodological consistency with classical coherence formulations and static embedding approaches, which operate on unordered lists of topic words. As a result, the reported findings for contextualized models should be interpreted as characterizing their behavior under a context-free usage regime, rather than as a comprehensive assessment of their full contextual modeling capabilities. For the aggregation phase, we compute the mean, and for the confirmation phase, we consider cosine similarity and dot product.

We leverage standard **datasets** widely used in the literature. Previous research, such as that by Aletras, Röder, and Ramrakhiyani [15, 22, 23], has relied on these datasets due to their accessibility and the depth of the available annotations. In particular, thanks to the generosity of the authors of [22], we have topics generated using the LDA model, accompanied by quality annotations made by human reviewers. These topics were derived from texts from three different collections, described below:

- **New York Times:** This dataset includes 47,229 news articles published between May and December 2010, extracted from the GigaWord corpus. In the original experiment, 200 topics were generated, of which 100 were selected for evaluation [22].
- **The 20 News Group Data Collection:** This consists of 20,000 emails organized into 20 thematic categories. Each category contains approximately 1,000 documents. In the original experiment, 100 topics were generated from this dataset [22].
- **Genomics:** This dataset includes 30,000 scientific articles published in 49 journals indexed in MEDLINE. Originally, these texts were used in the TREC Genomics Track. In the topic analysis study, 100 topics were identified [22].

We, therefore, have access to 100 topics per collection, and topics are represented by lists of 10 keywords per topic. The quality of each topic was evaluated by a group of human reviewers, ranging from 24 to 26 participants. Each reviewer was shown the 10 top words representing a topic and was asked to assign a coherence value from 1 to 3 (1 "Useless", 2 "Average", and 3 "Useful"). Each participant evaluated up to 100 topics. It is important to note that not all topics were evaluated by the same number of reviewers. To ensure reliable assessment, the creators of these topic collections introduced trap topics with obvious answers, and all annotations from individuals who failed to respond correctly to these cases were removed.

The preprocessing applied in our experiments differs across coherence metrics as a direct consequence of their underlying computational requirements. Classical metrics compute co-occurrence statistics from a reference corpus, requiring explicit corpus-level preprocessing to obtain reliable probability estimates. All documents in the reference corpus are lowercased, punctuation and numeric characters are removed, and the remaining text is tokenized. English stopwords are filtered out, and tokens are lemmatized using WordNet. The processed texts are then used to construct the dictionary and corpus representations from which co-occurrence statistics are derived. When computing coherence scores, only topic words present in the reference corpus vocabulary are retained, while out-of-vocabulary words are discarded.

In contrast, embedding-based metrics operate on pretrained vector representations whose parameters encode distributional information acquired from large-scale corpora during pre-training. Consequently, no corpus-level preprocessing or co-occurrence computation is required at evaluation time. For static embedding models (GloVe, Word2Vec, and FastText), each topic word is directly mapped to its corresponding pretrained embedding. For transformer-based models (BERT, RoBERTa, ALBERT, and MPNet), standard tokenizer-based preprocessing is applied: each word is processed independently, representations are derived from input (token) embeddings, and when

a word is split into multiple subword units during tokenization, the resulting token embeddings are averaged to produce a single word-level representation.

To obtain co-occurrence and related statistics, classical coherence metrics use either the same corpus where topics were generated (as in UMass) or samples from Wikipedia (as in CV, UCI, or NPMI). For a more complete comparison, we experimented with both alternatives: estimates from the same corpus and estimates from Wikipedia. Every classical coherence metric was tested following both variants, and our performance report presents separate results for each variant, allowing us to understand the effect of the reference corpus on every classical metric.

Following standard practice, the evaluation is performed by calculating the **Spearman correlation** between the quality estimates generated by each metric and the human quality judgments. This correlation measures the relationship between the values assigned by the metric and the values provided by the human evaluators, assessing how closely both estimates are. Specifically, the value of the automated coherence metric is calculated for all topics, the corresponding ground truth values for the topics (human evaluation) are taken from the test collection, and the Spearman coefficient between both lists of scores is calculated.

Summing up, the **experimental procedure** includes the following steps:

1. Generation of coherence scores for each topic using the metrics CV, UMass, UCI, NPMI and embedding-based metrics.
2. Calculation of the Spearman correlation coefficient and corresponding p-value between the coherence scores of each metric and the human judgments for each dataset.
3. Comparative analysis of the performance of the metrics on the three datasets: 20NG, Genomics, and NYT.

7 Experimental Results

The evaluation results (see Table 2) are organized into four groups: (i) classical metrics whose probability estimates were computed on the same corpus in which the topics were generated, (ii) classical metrics whose probability estimates were computed on the Wikipedia, (iii) the Tbuckets method[1], and (iv) embedding-based models.

From the analysis of the **traditional metrics results**, we have identified the following key findings:

- **Influence of the reference corpus**: In general, classical metrics perform considerably better when their estimates are obtained from Wikipedia. This suggests that Wikipedia has a superior estimation capacity for evaluating topic coherence, providing a more stable representation of word co-occurrence. Note that the topic analysis corpora are generally smaller and more focused than Wikipedia and, thus, the superiority of a classical coherence method should not be concluded based solely on the corpus used to guide its estimations.

[1]The results of Tbuckets, due to the difficulties in reproducing this method, have been taken directly from [15].

metric			20NG	NYT	Geno	Avg.
classical models						
co-occurrence computed from the same corpus						
NPMI			0.473‡	0.786‡	0.485‡	0.581
UMASS			0.544‡	0.648‡	0.409‡	0.534
CV			0.502‡	0.562‡	0.475‡	0.513
UCI			0.487‡	0.758‡	0.466‡	0.570
co-occurrence estimated from Wikipedia						
NPMI			0.766‡	0.777‡	0.663‡	0.735
UMASS			0.662‡	0.496‡	0.510‡	0.556
CV			0.730‡	0.777‡	0.660‡	0.722
UCI			0.731‡	0.753‡	0.653‡	0.712
Ramrakhiyani et al [15]						
Tbuckets			**0.870**	**0.819**	**0.729**	**0.806**
embedding models						
	sim	agg				
BERT	m_{cos}	σ_a	−0.365‡	−0.041	0.014	−0.131
RoBERTa	m_{cos}	σ_a	0.040	0.432‡	0.099	0.190
ALBERT	m_{cos}	σ_a	0.414‡	0.623‡	0.322‡	0.453
MPNet	m_{cos}	σ_a	0.369‡	0.721‡	0.274	0.455
FastText	m_{cos}	σ_a	0.701‡	0.612‡	0.670‡	0.661
Word2Vec	m_{cos}	σ_a	0.523‡	0.589‡	0.620‡	0.578
GloVe	m_{cos}	σ_a	0.825‡	0.755‡	**0.745**‡	0.775
BERT	m_{dot}	σ_a	−0.208†	−0.111	0.242†	−0.026
RoBERTa	m_{dot}	σ_a	0.551‡	0.576‡	0.012	0.380
ALBERT	m_{dot}	σ_a	0.578‡	0.520‡	0.370	0.489
MPNet	m_{dot}	σ_a	0.592‡	0.733‡	0.454	0.593
FastText	m_{dot}	σ_a	0.585‡	0.510‡	0.580‡	0.559
Word2Vec	m_{dot}	σ_a	0.503‡	0.623‡	0.646‡	0.591
GloVe	m_{dot}	σ_a	**0.832**‡	**0.819**‡	0.741‡	**0.797**

Table 2 Spearman correlation results between automated coherence metrics and quality judgments made by humans. For the embedding models, the second column reports the similarity method and the third column specifies the aggregation method. In the datasets' columns, the correlations that are statistically significant according to a two-sided Spearman rank correlation test ($p < 0.05$) are marked with the symbol †, while those with $p < 0.01$ are marked with the symbol ‡.

- **Variability of the performance of classical metrics**: Significant inconsistencies are observed in the classical metrics depending on the evaluated corpus:
  - **NPMI** is the best performer on NYT (same corpus estimates, 0.786 correlation), but its performance is low on 20NG and Geno. Still, NPMI achieves the best average values (0.581 for the same corpus and 0.735 for Wikipedia).
  - **UMASS** yields relatively poor performance across all collections, with low average scores on both reference corpora.
  - **UCI** performs poorly, particularly on Geno, suggesting a lack of robustness in certain domains.

– **CV** achieves competitive values in some collections, but its mean performance is lower than that of NPMI.

In conclusion, classical metrics show notable performance variations, and none of them seems to be a reliable option for evaluating topic coherence across different domains.

Regarding **Tbuckets**, it achieves the best overall performance with an average of 0.806, outperforming all classical metrics for all collections. It stands out on 20NG and NYT, with values of 0.870 and 0.819, while its performance on Geno is slightly lower (0.729). These results indicate that this method is highly effective for evaluating topic coherence.

The **embedding-based methods** also show a large variance in performance. Some of these pre-trained representations are clearly ineffective at capturing the coherence of the topics' top words. The most remarkable models are GloVe and, to some extent, FastText (but FastText loses some ground with the dot product similarity). Among embedding methods, GloVe achieves the highest mean performance with cosine similarity (0.775) and with dot product (0.797); and its effectiveness is consistent across all collections (all Glove variants produce correlations higher than 0.7). In contrast, BERT (and, to some extent, RoBERTa) show low and inconsistent values. ALBERT and MPNET fall somewhere in between, yielding some competitive scores, but these models are still far from the performance of Glove.

Among the embedding-based models, GloVe would thus be the best choice for measuring topic coherence, with a performance comparable to Tbuckets and clearly surpassing all classical coherence metrics. GloVe exploits co-occurrence at a global level within a text corpus, thus capturing word relationships more subtly than Word2Vec, which trains based on local contexts. While Word2Vec uses context-window techniques (CBOW or Skip-gram) to learn word relationships, GloVe uses a co-occurrence matrix that represents the probability of two words appearing together in a corpus. This allows GloVe to capture general word associations. Additionally, as reported in Table 1, Tbuckets stands on GloVe as its base model. As a consequence, our experiments reinforce GloVe's position as a reference model for supporting the estimation of topic word coherence.

On the other hand, one of the main advantages of contextual models like BERT (or models from this family) lies in their ability to disambiguate words based on nearby words. Our estimates of topic coherence treat each top word as an independent unit. Therefore, the input to the embedding models lacks the context that BERT, RoBERTa, or related models could exploit. This explains their suboptimal performance compared to GloVe.

Finally, Figure 4 graphically depicts the performance of each embedding model with cosine similarity and dot product. Some models (e.g., FastText) work better with cosine similarity, while other models (e.g., BERT) seem to be better suited for dot product. The best performing model, GloVe, performs well with both alternatives.

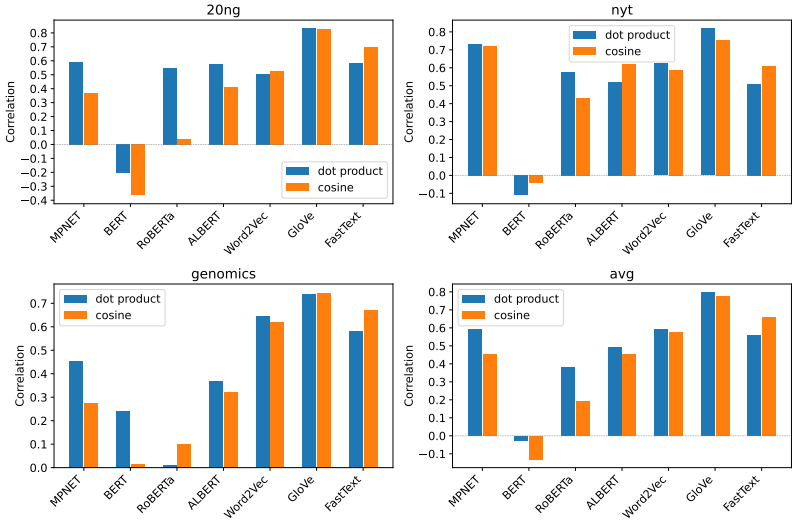


Fig. 4 Comparison of embedding models with different similarity metrics (cosine and dot product)

7.1 Ablation Analysis

Motivated by the poor performance of some transformer-based variants, we conducted an ablation study to investigate how different layers of these models capture information that might be relevant for estimating topic coherence. Although we focus on isolated words without sentential context, the transformer's intermediate layers apply a series of learned non-linear transformations that may encode linguistic regularities beyond surface lexical information. Analyzing representations across layers allows us to examine whether syntactic, semantic, or other abstract properties emerge even when the input consists of a single word. This comparison helps disentangle purely lexical information from higher-level features captured by the model's internal hierarchy. We selected BERT (the worst-performing transformer model) and MPNet (the best-performing transformer model) for this analysis.

We extracted representations from the input embedding layer and from all 12 subsequent transformer layers, and computed coherence scores using each layer's representation independently. Figure 5 shows the Spearman correlation between the resulting coherence estimates and human judgments across the three datasets. In these plots, the leftmost points correspond to the input embeddings, while the rightmost points correspond to the upper layers of the transformer.

The graphs show systematic differences in the behavior of the two models. MPNet achieves strong performance with the input embeddings ($\rho \approx 0.35$–0.70 across datasets) and maintains relatively stable correlations through the first two layers, after

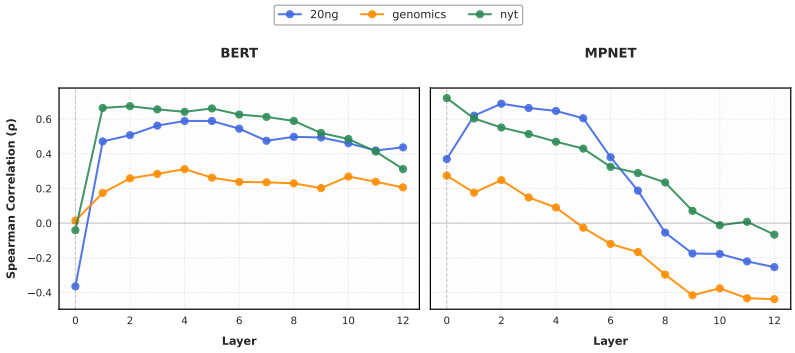


Fig. 5 Extracting word embeddings from different transformer layers. Effect on topic coherence performance

which performance steadily degrades. This pattern suggests that MPNet's early layers effectively capture the semantic relationships between individual words that are relevant for coherence evaluation, while deeper layers increasingly encode additional information that may be less useful.

In contrast, BERT shows poor performance at the input embedding layer ($\rho \approx -0.40$ to 0.00), but correlation improves dramatically after the first transformer layer, reaching peak performance around layers 2–5 ($\rho \approx 0.25$–0.65). After reaching this peak performance, BERT exhibits a degradation pattern similar to that of MPNet. This substantial improvement from the embedding layer to the first transformer layer can be attributed to BERT's subword tokenization. Individual subword tokens often carry only partial semantic information; however, after the first layer's contextual transformation, their representations become more integrated, so that averaging them yields more coherent word-level embeddings. This enrichment effect saturates early, after which deeper layers exhibit a degradation trend similar to that observed for MPNet. However, the decline is considerably more gradual, with correlation values consistently remaining above 0.2.

These findings have important implications: i) the optimal layer for extracting representations for estimating topic coherence varies across model architectures, ii) deeper contextual representations are not necessarily better for word-level semantic similarity tasks, and iii) the poor performance of some transformers in experiments could be partially mitigated by careful layer selection. In any case, even when selecting representations from the empirically optimal layer, performance still falls short of that achieved by static embeddings such as GloVe.

7.2 Error Analysis

To better understand how coherence metrics fail, we now analyze two types of error patterns: overestimation (cases in which models rate coherence higher than human

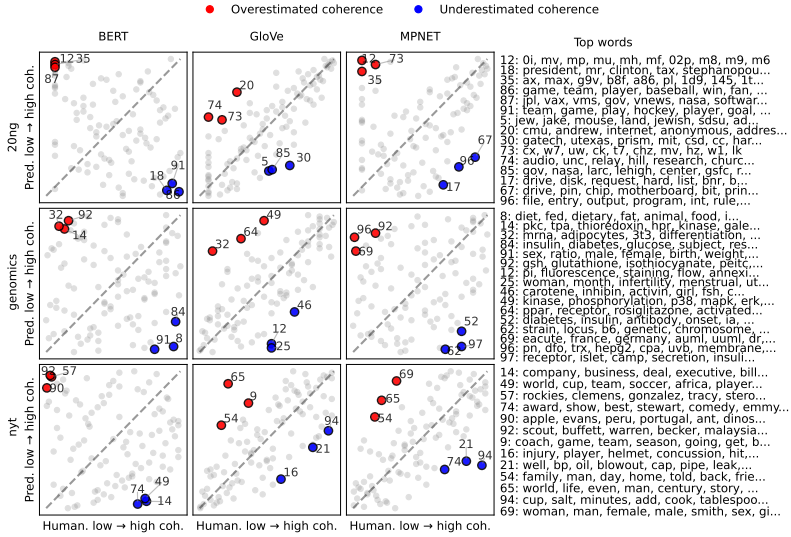


Fig. 6 The blue (red) topics represent topics in which human evaluators perceived high (low) coherence, but the models estimated low (high) coherence.

judgments) and underestimation (cases in which models rate coherence lower than human judgments). Given the available set of topics, we can analyze how model-estimated coherence differs from human-perceived coherence by comparing their respective topic rankings and examining, for instance, whether the topics rated as most coherent by humans are also assigned high coherence by automatic metrics.

In the resulting scatter plots (Figure 6), the x-axis corresponds to human rankings and the y-axis corresponds to model rankings. Topics near the diagonal indicate accurate relative coherence estimates by the model. Topics below the diagonal denote underestimation errors, while topics above the diagonal denote overestimation errors.

GloVe exhibits the tightest clustering around the diagonal across all datasets, reflecting its strong overall alignment with human rankings. In contrast, BERT exhibits larger deviations, with points distributed farther from the diagonal. MPNet shows an intermediate level of dispersion.

A closer inspection of BERT's overestimation errors shows that they frequently correspond to highly cryptic or specialized topics. In 20NG, for example, several overestimated topics consist primarily of abbreviated codes or technical acronyms (e.g., *0i, mv, mp, mu*, or *jpl, vax, vms, nasa*). Similarly, in Genomics, overestimated topics contain specialized biochemical terminology (e.g., *pkc, tpa, thioredoxin*, *mrna, adipocytes, 3t3*). While such word lists may appear incoherent to non-expert annotators, embedding models trained on large corpora can capture distributional regularities

among these co-occurring terms. Consequently, these overestimation errors may reflect latent semantic structure that is less accessible to human evaluators without domain expertise.

In contrast, underestimation errors are more problematic. The most severely underestimated topics consist of common, everyday vocabulary forming clearly interpretable themes, such as politics or sports (e.g., *president, clinton, tax*; *game, team, player*; *company, business, deal*). In these cases, semantic relatedness is readily apparent to human annotators. Such failures suggest that contextualized representations and subword tokenization—while powerful for many NLP tasks—may not optimally capture the type of straightforward word-level similarity required for topic coherence evaluation.

Related to this, our future work will focus on extending the proposed evaluation framework to better exploit the representational capabilities of contextualized language models while preserving compatibility with topic coherence formulations based on top-word lists. In particular, we plan to investigate template-based pseudo-contexts [41, 42], where isolated topic words are embedded within simple, generic sentence templates. This provides minimal yet controlled contextual information without introducing topic-specific bias. We also intend to explore pairwise contextualization approaches [43], in which pairs of topic words are jointly embedded, allowing transformer models to capture relational semantics while remaining aligned with the pairwise comparison paradigm commonly adopted in coherence metrics. Finally, we will analyze multi-layer pooling strategies for transformer-based models [44, 45], such as averaging representations from higher layers (e.g., layers 9–12), which have been shown to encode richer semantic information than input embeddings alone. Together, these extensions aim to provide a more faithful assessment of contextualized models for topic coherence evaluation and to clarify the extent to which their contextual modeling strengths can improve coherence estimation beyond our current usage.

8 Conclusions

In this study, we have compared classical topic coherence metrics with embedding-based approaches. Modern methods, based on dense vector representations of words, can outperform traditional metrics in many settings, although their performance depends on model architecture and layer selection. While classical metrics can provide valuable information, their performance is generally suboptimal, leaving clear room for improvement; a simple method based on static embeddings such as GloVe consistently surpasses them. Among embedding-based approaches, GloVe in particular achieves strong correlations with human judgments, especially for topics containing common vocabulary, demonstrating its robustness in capturing word-level semantic similarity.

Our ablation analysis revealed that transformer-based models such as BERT and MPNet exhibit layer-dependent behavior. Early layers often capture semantic relationships relevant to coherence more effectively than deeper layers, whereas representations from the final layers tend to encode richer contextual information that may not align with human assessments of word-level coherence. Specifically, BERT shows a sharp

improvement in layers 2–5 from initially poor embeddings, followed by gradual degradation, while MPNet maintains relatively high correlations in early layers before a steady decline.

Error analysis further highlighted systematic patterns: BERT tends to overestimate the coherence of cryptic or domain-specific topics (e.g., abbreviations or technical acronyms) and underestimate everyday topics that are readily interpretable by humans. These findings imply that, while transformer-based models have the potential to capture richer relational semantics, their default usage for topic coherence—especially using full forward-pass representations without layer selection or controlled contextualization—may not fully leverage their capabilities. Static embeddings remain a robust baseline for tasks requiring straightforward word-level similarity, whereas contextualized models may benefit from strategies such as template-based pseudo-contexts, pairwise contextualization of topic words, or multi-layer pooling to better align with human coherence judgments.

The conclusions drawn in this study are based on topic representations produced by classical LDA models and evaluated on widely used benchmark collections (20NG, NYT, and Genomics). Accordingly, findings such as the relatively strong performance of GloVe embeddings, the layer-dependent behavior of transformer-based models, and the asymmetric error patterns observed for BERT should be interpreted within this experimental scope. Extending the evaluation to neural topic models, alternative topic representations, and additional domains is necessary to assess the generality of these observations and to determine whether similar patterns hold across different settings.

Future research could further investigate methods to better exploit transformer-based models for topic coherence, including hybrid approaches with static embeddings, systematic ablations, and analysis of embedding dimensionality and similarity functions, while preserving comparability with top-word-based metrics.

9 Declarations

Competing Interests. The authors report there are no competing interests to declare.

Funding. The first and third authors thank the financial support supplied by the Agencia Estatal de Investigación (Spain) (PID2022-137061OB-C22; MCIN/AEI/10.13039/501100011033, Plan de Recuperación, Transformación y Resiliencia, Unión Europea-Next Generation EU), Consellería de Cultura, Educación, Formación Profesional e Universidades (Centro de investigación de Galicia accreditation 2024-2027 ED431G-2023/04 and Reference Competitive Group accreditation 2022-2025, ED431C 2022/19) and the European Union (European Regional Development Fund - ERDF).

The second author thanks the financial support supplied from projects: PID2022-137061OB-C21 (MCIN/AEI/10 .13039/501100011033/, Ministerio de Ciencia e Innovación, ERDF A way of making Europe, by the European Union); Consellería de Educación, Universidade e Formación Profesional, Spain (accreditations 2019–2022

ED431G/01 and GPC ED431B 2022/33) and the European Regional Development Fund, which acknowledges the CITIC Research Center.

Data. The datasets (20NG, NYT, Geno) used in this study are publicly available. The LDA topics and human annotations were obtained by contacting the authors of [22].

Code Availability. The code to reproduce the experiments is available at Github.

Authors' contribution statements. MC: conceptualization, methodology, investigation, formal analysis, and writing—original draft. JP and DEL: conceptualization, methodology, investigation, formal analysis, writing—review, editing, supervision, and project administration.

References

[1] Blei, D.M.: Probabilistic topic models. Communications of the ACM **55**(4), 77–84 (2012)

[2] Boyd-Graber, J., Hu, Y., Mimno, D., *et al.*: Applications of topic models. Foundations and Trends® in Information Retrieval **11**(2-3), 143–296 (2017)

[3] Alghamdi, R., Alfalqi, K.: A survey of topic modeling in text mining. Int. J. Adv. Comput. Sci. Appl.(IJACSA) **6**(1) (2015)

[4] Xu, W., Hirami, K., Eguchi, K.: Self-supervised learning for neural topic models with variance–invariance–covariance regularization. Knowledge and Information Systems, 1–19 (2025)

[5] Hoyle, A., Goel, P., Hian-Cheong, A., Peskov, D., Boyd-Graber, J., Resnik, P.: Is automated topic model evaluation broken? the incoherence of coherence. Advances in neural information processing systems **34**, 2018–2033 (2021)

[6] Rosner, F., Hinneburg, A., Röder, M., Nettling, M., Both, A.: Evaluating topic coherence measures. arXiv (2014). arXiv:1403.6397

[7] Rahimi, H., Hoover, J.L., Mimno, D., Naacke, H., Constantin, C., Amann, B.: Contextualized topic coherence metrics. arXiv preprint arXiv:2305.14587 (2023)

[8] Zhou, R., Wang, F., Liu, C., Lu, X.: Sparse dynamic topic model with topic birth and death over time. Knowledge and Information Systems, 1–23 (2025)

[9] Chang, C.-H., Hwang, S.-Y.: A word embedding-based approach to cross-lingual topic modeling. Knowledge and Information Systems **63**(6), 1529–1555 (2021)

[10] Gao, W., Peng, M., Wang, H., Zhang, Y., Xie, Q., Tian, G.: Incorporating word embeddings into topic modeling of short text. Knowledge and Information Systems **61**, 1123–1145 (2019)

[11] Du, C., Yao, K., Zhu, H., Wang, D., Zhuang, F., Xiong, H.: Mining technology trends in scientific publications: a graph propagated neural topic modeling approach. Knowledge and Information Systems **66**(5), 3085–3114 (2024)

[12] Nikolenko, S.I.: Topic quality metrics based on distributed word representations. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '16, pp. 1029–1032. Association for Computing Machinery, New York, NY, USA (2016)

[13] Fang, A., Macdonald, C., Ounis, I., Habel, P.: Using word embedding to evaluate the coherence of topics from twitter data. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1057–1060 (2016)

[14] Belford, M., Greene, D.: Comparison of embedding techniques for topic modeling coherence measures. In: LDK (Posters), pp. 1–5 (2019)

[15] Ramrakhiyani, N., Pawar, S., Hingmire, S., Palshikar, G.: Measuring topic coherence through optimal word buckets. In: Lapata, M., Blunsom, P., Koller, A. (eds.) Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pp. 437–442. Association for Computational Linguistics, Valencia, Spain (2017)

[16] Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., Blei, D.: Reading tea leaves: How humans interpret topic models. Advances in neural information processing systems **22** (2009)

[17] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3**, 993–1022 (2003)

[18] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American society for information science **41**(6), 391–407 (1990)

[19] AlSumait, L., Barbará, D., Gentle, J., Domeniconi, C.: Topic significance ranking of LDA generative models. In: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2009, Bled, Slovenia, September 7-11, 2009, Proceedings, Part I 20, pp. 67–82 (2009). Springer

[20] Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 100–108 (2010)

[21] Mimno, D., Wallach, H., Talley, E., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 262–272 (2011)

[22] Aletras, N., Stevenson, M.: Evaluating topic coherence using distributional semantics. In: Koller, A., Erk, K. (eds.) Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers, pp. 13–22. Association for Computational Linguistics, Potsdam, Germany (2013). https://aclanthology.org/W13-0102

[23] Röder, M., Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, pp. 399–408 (2015)

[24] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. arXiv. arXiv:1301.3781 [cs] (2013)

[25] Pennington, J., Socher, R., Manning, C.: GloVe: Global Vectors for Word Representation. In: Moschitti, A., Pang, B., Daelemans, W. (eds.) Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543. Association for Computational Linguistics, Doha, Qatar (2014)

[26] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching Word Vectors with Subword Information. arXiv. arXiv:1607.04606 [cs] (2017)

[27] Doogan, C., Buntine, W.: Topic model or topic twaddle? re-evaluating demantic interpretability measures. In: North American Association for Computational Linguistics 2021, pp. 3824–3848 (2021). Association for Computational Linguistics (ACL)

[28] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Bidirectional encoder representations from transformers. arXiv preprint arXiv:1810.04805, 15 (2018)

[29] Liu, Y.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 **364** (2019)

[30] Lan, Z.: Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942 (2019)

[31] Röder, M., Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. WSDM '15, pp. 399–408. Association for Computing Machinery, New York, NY, USA (2015)

[32] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc., USA (2017)

[33] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv. arXiv:1907.11692 [cs] (2019)

[34] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. arXiv. arXiv:1909.11942 [cs] (2020)

[35] Song, K., Tan, X., Qin, T., Lu, J., Liu, T.-Y.: MPNET: Masked and permuted pre-training for language understanding. Advances in neural information processing systems **33**, 16857–16867 (2020)

[36] Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. Journal of machine learning research **3**(Feb), 1137–1155 (2003)

[37] Hinton, G.E.: Distributed representations (1984)

[38] Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. nature **323**(6088), 533–536 (1986)

[39] Elman, J.L.: Finding structure in time. Cognitive science **14**(2), 179–211 (1990)

[40] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems **32** (2019)

[41] Bianchi, F., Terragni, S., Hovy, D., Nozza, D., Fersini, E.: Cross-lingual contextualized topic models with zero-shot learning. In: Merlo, P., Tiedemann, J., Tsarfaty, R. (eds.) Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 1676–1683. Association for Computational Linguistics, Online (2021)

[42] Grootendorst, M.: BERTopic: Neural topic modeling with a class-based tf-idf procedure. arXiv preprint arXiv:2203.05794 (2022)

[43] Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3982–3992. Association for Computational Linguistics, Hong Kong, China (2019)

[44] Jawahar, G., Sagot, B., Seddah, D.: What does BERT learn about the structure of language? In: Korhonen, A., Traum, D., Màrquez, L. (eds.) Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3651–3657. Association for Computational Linguistics, Florence, Italy (2019)

[45] Tenney, I., Das, D., Pavlick, E.: BERT rediscovers the classical NLP pipeline. In: Korhonen, A., Traum, D., Màrquez, L. (eds.) Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4593–4601. Association for Computational Linguistics, Florence, Italy (2019)

CHAPTER 6

GENERAL DISCUSSION

The first two publications of this thesis –the clustering comparison (Publication I, Chapter 4.1) and the topic modeling study (Publication II, Chapter 4.2)– can be understood as successive attempts to extract meaningful thematic structure from mental health-related social media data. While the introduction presented the technical foundations of clustering algorithms and topic models, this section reflects on the empirical lessons learned from applying these methods and on the cross-cutting insights that emerged when the results of both studies are considered jointly.

6.1 Lessons from the clustering study

The clustering comparison in Publication I yielded modest quantitative results: the best-performing configuration –Gaussian Mixture Models with RoBERTa embeddings– achieved an Adjusted Rand Index of 0.180 and a Normalized Mutual Information of 0.200. These values, while statistically above chance, indicate that the clustering algorithms struggled to recover the known diagnostic categories (depressed vs. control users) from the text representations alone. This outcome, however, was not entirely unexpected. The distinction between users with and without depression is unlikely to manifest as a clean geometric separation in any embedding space, since users discuss a wide variety of topics regardless of their mental health status.

More instructive than the quantitative results was the methodological insight regarding evaluation metrics. The study revealed a systematic discrepancy between internal and external validation: configurations using raw Term Frequency vectors consistently achieved the highest Silhouette Coefficients and Calinski-Harabasz scores, yet performed worst on external metrics

such as ARI and NMI. The explanation lies in the geometric properties of the representation spaces. TF vectors are unnormalized, meaning that their magnitude grows with document length. Internal metrics –which measure cluster compactness and separation through distances in the feature space– were effectively measuring variation in text length rather than semantic content. By contrast, representations produced by transformer models (RoBERTa, BERT, GPT-2) and TF-IDF normalization constrain vectors to comparable magnitudes, allowing distance-based metrics to reflect genuine semantic similarity.

This finding carries a broader implication: **internal clustering metrics are not comparable across representation spaces with different geometric properties**. The Silhouette Coefficient of a clustering performed on TF vectors and the Silhouette Coefficient of a clustering performed on RoBERTa embeddings are computed under fundamentally different distance functions. Treating them as commensurable –as is commonly done in comparative studies– can lead to misleading conclusions. This is not merely a theoretical concern; in our experiments, naïve reliance on internal metrics would have led to the selection of the worst-performing representation method.

The clustering study also provided substantive insights into the mental health domain. Despite the weak overall separability, the best configurations revealed meaningful thematic clusters: one cluster concentrated texts about relationships and dating, another about video games, and others captured broader discussion topics. Notably, the cluster associated with relationship difficulties showed a higher proportion of posts from depressed users, consistent with psychological literature linking interpersonal problems to depression. The analysis also identified correlations between depression status and participation in specific subreddits (e.g., those related to substance use and ADHD), suggesting that topic preferences carry diagnostic signal even when direct text-based classification remains challenging.

6.2 The transition to topic modeling

The limitations of the clustering approach motivated a natural evolution toward topic modeling. This transition was not a change of direction but rather a deepening of the same research question. Both clustering and modern topic models such as BERTopic share a common computational backbone: text representation through embeddings, dimensionality reduction (typically via UMAP), and density-based or centroid-based grouping. The critical difference is what happens *after* the grouping step. Clustering stops at the partition; topic modeling adds

a description layer –in BERTopic's case, the class-based TF-IDF (c-TF-IDF) procedure– that extracts representative terms for each group, transforming opaque clusters into interpretable topics. This description layer also introduces a family of coherence metrics (NPMI, C_V, UCI, UMass) that allow hyperparameter optimization to be guided by the descriptive quality of the discovered topics rather than by the geometric properties of the underlying groupings –directly addressing the weakness exposed in the clustering study.

This added interpretability, and the promise of content-oriented evaluation, proved transformative. Where the clustering study could only characterize clusters post-hoc by examining their constituent documents, BERTopic produced explicit topic descriptors that could be immediately assessed for thematic coherence. The topic modeling comparison in Publication II demonstrated that BERTopic significantly outperformed both LDA and TopClus in human evaluation: annotators rated BERTopic topics at an average coherence of 2.490 on a 3-point Likert scale, compared to 1.138 for LDA and 1.606 for TopClus. The inter-annotator agreement (Fleiss' $\kappa \approx 0.50$) was moderate, indicating that topic quality judgments, while subjective, were sufficiently consistent to support meaningful comparison.

BERTopic's advantage on social media data stems primarily from its use of pre-trained sentence transformers, which are inherently more robust to the lexical variability and noise characteristic of informal online text than LDA's bag-of-words representations. It should be noted, however, that LDA was evaluated with default hyperparameters –a significant caveat, since LDA's performance is known to be highly sensitive to the choice of priors and number of topics. This decision, while ensuring a fair out-of-the-box comparison, may substantially underestimate LDA's potential when carefully tuned for this domain. The extent to which BERTopic's superiority would hold against a properly optimized LDA remains an open question.

Beyond the model comparison, Publication II introduced a clinical relevance dimension that was absent from the clustering study. By applying BDI-II (Beck Depression Inventory) scoring through large language model prompting, the study demonstrated that the discovered topics were not merely statistically coherent but also clinically meaningful. Topics related to self-harm, relationship difficulties, and weight loss showed strong associations with depression indicators, providing evidence that topic modeling can bridge the gap between computational text analysis and clinical interpretation. The temporal analysis further revealed a sharp increase in depression-related topics during the 2020–2021 pandemic period, demonstrating the sensitivity of the approach to real-world events.

6.3 The evaluation problem: a recurring theme

Perhaps the most significant cross-cutting finding from these two studies is the pervasiveness of the evaluation problem. In the clustering study, internal metrics proved unreliable for cross-space comparison due to their sensitivity to the geometry of the representation space. In the topic modeling study, the evaluation problem resurfaced in a different form: the automatic coherence metrics commonly used to assess topic quality showed negligible correlation with human judgments. Specifically, NPMI achieved a Spearman correlation of only 0.112 with annotator ratings, while UMass (0.001), UCI (0.04), and C_V (0.05) were essentially uncorrelated.

All four classical coherence metrics share a fundamental dependency on word co-occurrence statistics derived from a reference corpus. When the reference corpus adequately represents the vocabulary and domain of the topic model, these statistics are informative. But when vocabularies diverge or the domain shifts –as is the case with informal mental health text on social media– the metrics become unreliable. In our experiments, topics that human annotators rated as highly coherent sometimes received negative NPMI scores, while topics rated as incoherent received positive scores. This disconnection, while documented in the broader literature [35, 27], was particularly severe in the mental health domain.

This observation –that the metrics commonly used to evaluate unsupervised text analysis carry hidden dependencies on the reference data– constitutes the central methodological insight connecting the first three publications. It motivated the systematic investigation of alternative evaluation approaches in Publication III, which we discuss in the next section.

6.4 Towards Robust Coherence Metrics

The evaluation challenges identified in the clustering and topic modeling studies –the geometry-dependence of internal clustering metrics and the corpus-dependence of classical coherence metrics– share a common structure: in both cases, the metric's reliability is contingent on properties of the representation or reference data that are often invisible to the practitioner. Publication III (Chapter 5.1) addressed the coherence metric problem directly by conducting a systematic evaluation of embedding-based alternatives to classical co-occurrence metrics.

6.4.1 The corpus dependency of classical metrics

Before discussing the embedding-based alternatives, it is worth quantifying the extent of the corpus dependency problem. In Publication III, we evaluated classical coherence metrics (NPMI, C_V, UCI, and UMass) under two conditions: using the original training corpus for co-occurrence estimation, and using Wikipedia as an external reference corpus. The results revealed a striking difference. NPMI's average Spearman correlation with human judgments improved from 0.581 to 0.735 when switching from the original corpus to Wikipedia; C_V showed an even more dramatic improvement, from 0.513 to 0.722. These results confirm that classical metrics are not measuring an intrinsic property of the topics but rather a joint property of the topics *and* the reference corpus. When the reference corpus is large, general-purpose, and lexically rich (as Wikipedia is), co-occurrence statistics are more informative and the metrics become more reliable. When the reference corpus is small, domain-specific, or noisy, the metrics degrade.

This finding has direct implications for the topic modeling study in Publication II, where the negligible correlation between automatic metrics and human judgments can now be partially explained. The eRisk dataset used in that study consists of Reddit posts about mental health –a corpus that is far from the encyclopedic, well-edited text of Wikipedia. The co-occurrence patterns in this corpus reflect the informal, repetitive, and emotionally charged language of online support communities, which diverges substantially from the patterns captured in general-purpose reference corpora. It is therefore unsurprising that coherence metrics computed on this data failed to align with human assessments.

6.4.2 Embedding-based coherence: eliminating the reference corpus

The central contribution of Publication III was to evaluate whether pre-trained word embeddings could serve as an alternative basis for coherence measurement, thereby eliminating the dependency on a reference corpus entirely. The rationale is straightforward: word embeddings encode semantic relationships learned during pre-training on large, diverse corpora. By measuring the pairwise similarity between the embedding vectors of a topic's top words, one can obtain a coherence score that reflects semantic relatedness without requiring any corpus-specific co-occurrence computation.

We evaluated seven embedding models spanning two families: static embeddings (Word2Vec,

GloVe, FastText) and contextual embeddings from transformer architectures (BERT, RoBERTa, ALBERT, MPNet). The evaluation was conducted on three standard datasets with human coherence annotations (20 Newsgroups, New York Times, Genomics), each containing 100 LDA topics rated by 24–26 human annotators on a 3-point scale.

The results revealed a clear hierarchy. Among all embedding models, GloVe achieved the highest average correlation with human judgments: 0.775 using cosine similarity and 0.797 using dot product. This performance not only exceeded all other embedding models but also surpassed the best classical metric (NPMI with Wikipedia, at 0.735). Crucially, GloVe's performance was consistent across all three datasets, with correlations exceeding 0.74 in every case –a level of cross-domain robustness that no classical metric matched.

The strong performance of GloVe can be attributed to its training objective, which explicitly factorizes the global word-word co-occurrence matrix. This means that GloVe embeddings encode precisely the kind of statistical associations that coherence metrics are designed to capture, but in a pre-computed, transferable form that does not require a reference corpus at evaluation time. In effect, GloVe compresses the co-occurrence statistics of its training corpus into a dense vector space, and the similarity between word vectors in this space serves as a proxy for co-occurrence-based coherence.

6.4.3 The surprising failure of contextual embeddings

One of the most unexpected findings of Publication III was the poor performance of contextual embeddings –particularly BERT and RoBERTa– for coherence measurement. BERT achieved a negative average correlation (-0.131 with cosine similarity), meaning that its similarity scores were inversely related to human coherence judgments. RoBERTa performed only marginally better (0.190). This result is striking given that BERT and RoBERTa are generally considered more powerful language models than GloVe.

The explanation lies in the nature of the coherence evaluation task. Topic coherence is assessed by presenting a list of isolated words (the top-N words of a topic) and measuring their pairwise similarity. Contextual embedding models are designed to produce representations that are sensitive to the surrounding linguistic context: the embedding of a word changes depending on the sentence in which it appears. When words are presented in isolation –as they are in coherence evaluation– contextual models lack the sentence-level context they were designed to exploit. Their input embeddings, which are the only layer available for isolated

words, have not been refined by the attention mechanism and therefore capture only shallow lexical information.

An ablation study across transformer layers confirmed this interpretation. BERT's coherence performance improved dramatically when using intermediate layers (layers 2–5), where the model has begun to integrate positional and structural information but has not yet specialized for task-specific contextual disambiguation. However, even at these optimal layers, BERT did not consistently outperform GloVe. MPNet showed a complementary pattern: strong performance at the input layer that degraded in deeper layers, suggesting that its pre-training objective preserved more useful word-level information in the initial representations.

This finding carries a practical recommendation: **for word-level semantic tasks such as coherence evaluation, static embeddings remain preferable to contextual models**. The additional complexity and computational cost of transformer-based models does not translate into better performance when the input consists of isolated words rather than contextualized sentences. This insight connects back to the clustering study (Publication I), where RoBERTa embeddings proved superior for document-level representation –a task where full sentences provide the context that transformers require– while the word-level coherence task favors the simpler, context-independent representations of GloVe.

6.4.4 Implications for topic model evaluation

The results of Publication III suggest a practical framework for topic model evaluation that addresses the limitations identified in Publications I and II. Rather than relying on classical coherence metrics with their hidden corpus dependencies, researchers can use GloVe-based coherence as a more robust alternative. This metric is pre-computed (requiring no reference corpus at evaluation time), cross-domain robust (consistent performance across journalism, newsgroups, and biomedical text), and better aligned with human judgments than any classical alternative tested. It is worth noting that TBuckets –a method based on SVD decomposition and integer linear programming– achieved slightly higher average correlation (0.806) than GloVe (0.797), suggesting that algebraic approaches to word grouping offer a complementary perspective that merits further investigation.

However, a significant gap must be acknowledged between the evaluation setting and the motivating application. Publication III evaluated its metrics on standard topic modeling benchmarks (20 Newsgroups, New York Times, Genomics) using topics generated by LDA –all well-edited, general-purpose text. The mental health social media domain that motivated

the investigation was not included. Recommending GloVe-based coherence for mental health applications on the basis of its performance on newswire and biomedical text involves an argumentative leap that the cross-domain robustness results encourage but do not fully justify. Constructing human-annotated topic coherence benchmarks on mental health social media data and evaluating embedding-based metrics directly on them is essential to close this gap.

6.5 Temporal Word Embeddings for Early Detection

While the first three publications focused on understanding the thematic structure of mental health text, Publication IV (Chapter 4.3) shifted toward a predictive goal: the early detection of psychological disorders from social media. The approach –tracking how individual users' language changes over time through temporal word embeddings– produced competitive results across 18 datasets and 4 disorders, but also revealed fundamental tensions in the pipeline design that merit careful discussion.

6.5.1 Temporal information as a diagnostic signal

The core insight motivating Publication IV is that mental health conditions do not appear instantaneously in language; rather, they manifest through gradual shifts in how individuals use words over time. A user transitioning toward a depressive episode may progressively shift their use of words like "sleep" or "feel" from neutral contexts to distress-related ones. By computing temporal word embeddings –separate embedding spaces for successive time periods of a user's posting history– and measuring the magnitude of word movements (deltas) between these temporal slices and a reference embedding, the framework captures these linguistic trajectories as quantitative features.

The results confirmed that temporal information carries genuine diagnostic value. Across the 18 eRisk datasets spanning depression, anorexia, self-harm, and gambling, the temporal embedding approach achieved competitive performance, particularly excelling on gambling detection (F1 = 0.84 with DCWE-SVM on eRisk 2023) and producing respectable results on self-harm and anorexia. These results demonstrate that word-level temporal dynamics –how the semantic context of specific words shifts over time– encode information relevant to mental health status that static, atemporal representations cannot capture.

Furthermore, the temporal perspective offers value beyond classification. The keyword analysis in Publication IV revealed that positive users (those with diagnosed conditions) exhibit

more pronounced and irregular word movements compared to control users. For gambling, this manifested as dramatic fluctuations in terms directly related to the disorder (“gambling”, “money”, “lost”). For depression, the patterns were more diffuse, spread across general emotional vocabulary (“feel”, “life”, “anxiety”) –consistent with the clustering study’s finding that depression does not manifest as a clean thematic separation. This temporal characterization can inform understanding of *when* and *how* disorders develop, not merely whether they are present.

6.5.2 The dense-to-sparse paradox

Despite these positive findings, the temporal embedding pipeline exhibits a fundamental architectural tension that limits its effectiveness. The computational cost of the approach is substantial, and understanding *why* it is so expensive reveals a deeper conceptual issue.

The cost arises because the framework must train a separate embedding model for each temporal slice of each user’s posting history. TWEC requires retraining a Word2Vec model for every time period, with no GPU acceleration, resulting in over three hours of computation per run. DCWE, while benefiting from GPU acceleration via PyTorch, must still process each user-period’s text through BERT to generate contextual embeddings. With multiple users, multiple time periods, and multiple datasets across four disorders, the aggregate computational burden is considerable.

This expense would be justified if the resulting features captured rich temporal information. However, the subsequent feature extraction step reduces the dense embeddings to scalar delta values –cosine distances between each word’s temporal embedding and its reference embedding– producing a sparse feature matrix where each column corresponds to a vocabulary word and each row to a user-period pair. This transformation discards most of the information that made the dense embeddings valuable. The direction of semantic movement is lost: a word that moves toward depression-related contexts and one that moves toward euphoria-related contexts produce the same delta magnitude. Multiple senses are collapsed: DCWE averages all contextual embeddings of a word within a time period, eliminating the sense distinctions that contextual models are designed to capture. And the resulting matrices are extremely sparse: TWEC produces approximately 50,000 word-features, DCWE approximately 9,000, but for any given user-period pair, only a small fraction will have been observed.

There is a more fundamental issue, however, that goes beyond the information bottleneck. Temporal word embedding methods such as TWEC and DCWE were originally designed to

study **diachronic semantic change** –how the meaning of words evolves over years or decades within a society. They are well suited for detecting sociological phenomena: the semantic shift of "cloud" from meteorology to computing, or the broadening of "tweet" from birdsong to social media. In these applications, the word genuinely acquires new meanings, and retraining embeddings for each time period captures this evolution effectively.

In mental health detection, however, the phenomenon of interest is fundamentally different. The word "sleep" does not change its meaning when a user transitions toward depression; what changes is the **emotional valence and personal context** in which the user employs it. A depressed user does not use "sleep" in a new sense; they use it in a distressed sense –"I can't sleep", "I sleep all day and still feel exhausted". The temporal embedding framework, by measuring how a word's distributional neighborhood shifts, captures this change only indirectly and noisily. It detects that the word has moved in the embedding space, but it cannot distinguish whether the movement reflects a genuine semantic change (irrelevant for mental health) or an emotional-contextual change (the actual signal of interest).

This mismatch between the tool and the task helps explain why the approach works better for gambling –where users genuinely introduce domain-specific vocabulary ("gambling", "bet", "casino") that is distributionally distinct– than for depression, where the relevant change is not *which* words are used but *how* emotionally loaded their usage becomes. For mental health detection, a more productive temporal dimension would track the emotional associations of words over time rather than their distributional neighborhoods. Sentiment-aware temporal representations, or architectures that explicitly model affect trajectories, would be more naturally aligned with the underlying phenomenon.

The comparison between TWEC and DCWE illustrates the trade-offs within the current paradigm. TWEC's larger vocabulary (50K features) captures more words but at higher computational cost and with greater sparsity. DCWE's filtered vocabulary (9K features) is more manageable, but the filtering may discard diagnostically relevant words. Neither approach resolves the fundamental issue: **dense embedding models designed for diachronic semantic change are being repurposed –at great computational expense– for a task whose signal lies in emotional-contextual shifts rather than meaning evolution**.

6.5.3 Vocabulary sensitivity in temporal features

The temporal embedding framework also inherits a version of the vocabulary dependency problem identified in the earlier publications. TWEC and DCWE construct their feature

spaces from the vocabulary observed during training. If test users express their distress using words that were absent or rare in the training set –a plausible scenario given the diversity of online self-expression– the corresponding word movements are simply not captured. The feature vector for such users will have zero values for the diagnostically relevant dimensions, effectively rendering the model blind to their linguistic signals.

This limitation echoes the vocabulary overlap problem identified for classical coherence metrics in Section 6.4: in both cases, the system's ability to evaluate or classify depends on the degree to which the target vocabulary aligns with the reference vocabulary. For coherence metrics, the reference is the co-occurrence corpus; for temporal embeddings, it is the training set's vocabulary. In both cases, users or topics that fall outside the expected vocabulary are systematically disadvantaged.

DCWE's vocabulary filtering partially addresses the computational cost of large vocabularies but exacerbates the coverage problem. By retaining only approximately 9,000 words (after removing stop words and rare terms), the framework gains computational efficiency at the potential cost of discarding words that, while infrequent in aggregate, may be highly diagnostic for specific individuals or disorders.

6.5.4 Disorder-dependent detectability

A striking finding of Publication IV was the dramatic variation in detection performance across disorders. Gambling showed the highest detectability (F1 = 0.84), while depression proved far more challenging (F1 = 0.35). The keyword analysis provides an explanation: gambling-related language is focused and distinctive ("gambling", "gamble", "money", "lost", "addiction"), with high correlation values between word usage and disorder status (top word correlation: 0.584). Depression-related language, by contrast, is diffuse and overlaps extensively with general emotional discourse (top word "feel" with correlation: 0.253).

This finding has implications beyond the specific framework used. It suggests that **the detectability of a mental health condition from text is partly determined by how linguistically distinctive the condition is** –how much its expression diverges from baseline language use. Conditions that produce focused, domain-specific vocabulary changes (gambling, eating disorders) are inherently more amenable to word-level detection methods than conditions like depression, which manifest through subtle shifts in the use of common, everyday words. This observation should inform the design of future detection systems: for linguistically diffuse conditions, methods that operate at the sentence or document level –capturing how common

words are combined rather than which individual words are used– may be more appropriate than word-level approaches.

6.5.5 Toward dense temporal architectures

The limitations identified above –the dense-to-sparse information bottleneck, the mismatch between diachronic semantic change and emotional-contextual shifts, the vocabulary sensitivity, and the disorder-dependent effectiveness– converge on a clear direction for future work.

Rather than repurposing models designed for diachronic analysis, future approaches should operate directly on dense temporal representations that are sensitive to affect rather than meaning change. Transformer-based architectures with temporal attention mechanisms could process sequences of dense document or sentence embeddings across time periods, preserving both the semantic richness and the directional information that the delta-based approach discards. Such architectures would also eliminate the dependency on a fixed vocabulary, since they would operate on continuous representations rather than discrete word identities.

More specifically, the emotional tracking dimension could be exploited through architectures that combine temporal sequence modeling with sentiment- or affect-aware representations. Rather than asking "how has the distributional neighborhood of this word changed?" (a diachronic question), such systems would ask "how has the emotional tone of this user's language evolved?" (a clinical question). This reframing –from word-level semantic change to user-level emotional trajectory– would align the computational methodology more directly with the clinical phenomenon it aims to detect, and could be combined with the temporal information that Publication IV demonstrated to be valuable.

The temporal embedding work thus serves as both a proof of concept –demonstrating that temporal linguistic dynamics carry diagnostic signal– and a motivation for approaches that can exploit this signal through representations more naturally suited to mental health detection.

6.6 Reproducible Experimentation Infrastructure

Publication V (Chapter 4.4) addresses a problem that is not specific to mental health analysis but that became acutely visible throughout the research program: the lack of standardized, efficient infrastructure for iterative scientific experimentation. LabChain, the framework presented in this publication, emerged directly from the practical frustrations encountered while conducting the research described in the preceding sections.

6.6.1 The reproducibility problem in practice

The computational methods employed across this thesis –clustering with multiple representations, topic modeling with several algorithms, temporal word embedding training, coherence metric evaluation– share a common workflow pattern: expensive intermediate computations (typically embedding generation) that are reused across multiple downstream analyses. Training TWEC embeddings on the eRisk datasets, for instance, requires over three hours of computation. When exploring five different classifiers on the same embeddings, a naïve implementation would repeat the embedding computation five times, wasting approximately twelve hours of compute.

This inefficiency, while significant in itself, reflects a deeper structural problem in how research code is organized. Throughout this thesis, we worked with models –TWEC, DCWE, and others– that operate far from the mainstream ecosystem of sentence-transformers, AutoModel, or adapter libraries. These are custom implementations with their own data loading conventions, configuration approaches, and output formats, often poorly documented and structurally incompatible with each other. This is not a deficiency unique to these projects; it is a systemic characteristic of research software, where the priority is publishing results rather than engineering reusable components. Each new model we integrated required substantial bespoke engineering effort, and the resulting experimental code was itself monolithic, fragile, and difficult for others to reproduce.

This experience revealed a gap in the existing tool landscape. Scikit-learn provides excellent abstractions for standard machine learning components but does not address caching or the integration of non-standard models. Workflow managers such as Snakemake and Luigi handle task dependencies but separate workflow specification from execution logic, requiring researchers to maintain both. Higher-level frameworks like Kedro and PyCaret offer useful abstractions but assume that components conform to their specific interfaces, which is rarely the case for novel research code.

6.6.2 From monolithic experiments to modular pipelines

LabChain addresses these limitations through two key design decisions. First, it adopts a pipes-and-filters architecture where each processing step (a *filter*) encapsulates a single transformation, and filters are composed into serializable pipelines. Public attributes (configuration parameters) are restricted to basic serializable types, ensuring that any pipeline can be fully

described as a JSON document. Private attributes handle non-serializable state (tokenizers, pre-trained models, fitted parameters). This separation means that a serialized pipeline is not merely a description –it is an executable specification that can be deserialized and run without additional code.

Second, LabChain implements content-addressable caching through deterministic hashing. Each filter's cache key is computed from its class name, public configuration, and –for trainable filters– the hash of its training data. When a filter processes input, the output hash combines the filter hash with the input hash, creating a provenance chain that uniquely identifies every intermediate result. If the same transformation is applied to the same data (even in a different experiment or by a different team member), the cached result is returned automatically. Crucially, this hashing is based on *what* the computation does (configuration + data), not on *how* or *where* it runs (source code version, library version, hardware). This makes caching robust to incidental environmental differences while remaining sensitive to meaningful changes in the experimental setup.

6.6.3 Impact on the research program

LabChain's contribution to this thesis was not merely conceptual; it had direct, measurable impact on the experimental work. The re-implementation of the temporal word embedding pipeline from Publication IV using LabChain's modular structure revealed a preprocessing bug –a failure to flatten hierarchical data before feeding it to TWEC– that had remained hidden in the original monolithic code. Fixing this bug, combined with the systematic exploration that LabChain's caching enabled, produced performance improvements of up to 192.3% on some tasks. This is a concrete illustration of how software engineering practices –modularization, explicit interfaces, automated caching– can amplify scientific discovery by making it easier to identify and correct errors, explore alternative configurations, and reproduce results reliably.

The computational savings were also substantial. For the mental health detection tasks alone (depression, anorexia, self-harm, gambling), LabChain's caching eliminated an estimated 48 hours of redundant computation, corresponding to approximately 2.50 kg of CO_2 emissions. While modest in absolute terms, these savings scale with the number of experiments and team members, and they represent time that researchers can redirect from waiting for computations to analyzing results.

LabChain is an evolving project –a refinement of work begun during the author's master's thesis, now published as open-source software and employed daily in production research

workflows at CiTIUS. It does not claim to solve all problems of research reproducibility, but it provides a practical starting point: a framework that makes the right thing (caching, modularization, serialization) the easy thing, reducing the friction that leads researchers to write monolithic, unreproducible experimental code.

6.7 Limitations

The studies discussed in this chapter share several limitations that should be acknowledged.

Data and language scope. All experiments on mental health data were conducted exclusively on English-language Reddit posts from the eRisk shared tasks. While Reddit provides a rich source of naturalistic mental health discourse, it represents a specific demographic and cultural context. The generalizability of our findings to other platforms (Twitter, clinical forums, non-English communities) remains untested. Furthermore, the diagnostic labels in eRisk are derived from self-reported diagnoses, which may not correspond to clinical assessments.

Algorithm coverage. The clustering comparison in Publication I evaluated three algorithms (K-Means, DBSCAN, GMM), excluding more recent approaches such as deep clustering methods and spectral clustering due to computational constraints. Similarly, the topic modeling comparison used default hyperparameters for LDA, which is known to be sensitive to tuning. While this choice ensured a fair comparison of out-of-the-box performance, it may underestimate LDA's potential when carefully optimized.

Coherence evaluation benchmarks. The embedding-based coherence metrics in Publication III were evaluated on established benchmarks (20 Newsgroups, New York Times, Genomics) using LDA-generated topics. These datasets, while widely used, consist of well-edited text that differs substantially from the noisy social media data that motivated our research. The crucial test –whether GloVe-based coherence outperforms classical metrics on mental health social media topics– has not yet been conducted. This gap between the motivation and evaluation domains is the most significant limitation of the current work.

Human evaluation scale. The human evaluation in Publication II used a simplified 3-point Likert scale to maximize inter-annotator agreement. While this design choice achieved its purpose ($\kappa \approx 0.50$), a finer-grained scale might have captured more nuanced distinctions in topic quality. The annotator pool, while adequate for the study's purposes, was limited in size and did not include clinical psychologists who might assess topic relevance differently.

Temporal embedding scope. The temporal word embedding framework in Publication IV

operates exclusively at the word level, capturing individual word movements but not sentence- or document-level temporal dynamics. The reduction from dense embeddings to scalar delta values discards directional semantic information. Additionally, the framework's feature space is defined by the training vocabulary, making it sensitive to vocabulary mismatch between training and test sets –a limitation that mirrors the corpus dependency problem identified for coherence metrics. The approach was not combined with ensemble methods or additional feature types, which likely accounts for the performance gap relative to state-of-the-art systems that leverage multiple complementary signals.

Bridging unsupervised and supervised evaluation. A broader limitation concerns the relationship between the unsupervised analyses presented here and clinical utility. While Publication II demonstrated that BERTopic topics align with BDI-II dimensions, this analysis relied on LLM-based scoring rather than expert clinical annotation. The extent to which automatically discovered topics are actionable for mental health professionals remains an open question that requires interdisciplinary collaboration.

LabChain adoption and scope. LabChain's content-addressable hashing deliberately excludes source code versions and library versions from cache keys unless explicitly encoded as public parameters. This design choice favors simplicity and cross-environment portability but means that changes in underlying library behavior (e.g., a new version of a tokenizer) will not automatically invalidate cached results. Furthermore, while LabChain has been validated within the CiTIUS research group, its adoption beyond this context remains limited, and the learning overhead for teams already invested in established frameworks such as Kedro or Snakemake may reduce its practical appeal.

6.8 Integration and Coherence

The five publications that constitute this thesis trace two interleaved threads: a progression from understanding data to building detection systems, and a recurring confrontation with the question of how we know our results are reliable.

The first thread begins with exploratory analysis. The clustering study (Publication I) demonstrated that unsupervised grouping of mental health texts is feasible but insufficient: the resulting clusters captured meaningful thematic structure but lacked built-in interpretability. Topic modeling (Publication II) addressed this gap by adding appropriate heuristics to the clustering pipeline, producing described, assessable topics. The temporal embedding study

(Publication IV) then extended the analysis from static snapshots to dynamic trajectories, demonstrating that temporal linguistic change carries a diagnostic signal for early detection. At each stage, the analysis deepened: from grouping, to explaining, to predicting.

The second thread –the evaluation problem– runs through the entire program. Internal clustering metrics proved geometry-dependent and misleading across representation spaces (Publication I). Classical coherence metrics proved corpus-dependent and nearly uncorrelated with human judgments on non-standard text (Publication II). Embedding-based coherence metrics offered a more robust alternative but shifted the dependency from reference corpora to pre-training data (Publication III). And temporal embeddings introduced a vocabulary sensitivity that echoes the same structural issue: features defined by a fixed vocabulary fail to capture users whose language falls outside it (Publication IV).

These are not independent problems; they are manifestations of a single underlying principle: **evaluation is not separable from representation**. Every metric, whether it measures cluster quality, topic coherence, or temporal word movement, operates within a representational context that shapes what it can and cannot measure. Recognizing these dependencies –and choosing evaluation strategies appropriate to the specific comparison being made– is essential for drawing valid conclusions from computational text analysis.

LabChain (Publication V) provides the practical infrastructure that made this multi-faceted investigation feasible. By automating the caching of expensive intermediate computations and enforcing modular pipeline design, it reduced the friction of iterative experimentation and enabled the systematic exploration required to uncover both the results and the methodological insights reported across the preceding publications. Its contribution is not to the science of mental health detection directly, but to the engineering of scientific practice: making reproducible, efficient experimentation the path of least resistance rather than an aspirational goal.

Taken together, these five publications offer a perspective on computational mental health analysis that is both technically grounded and methodologically self-aware. The empirical contributions –BERTopic's superiority on social media text, GloVe-based coherence as a robust evaluation metric, temporal embeddings as diagnostic features– are accompanied by a critical assessment of the tools and assumptions underlying those contributions. This combination of results and reflection constitutes the most durable contribution of the thesis.

CHAPTER 7

CONCLUSIONS

This chapter presents the principal conclusions of the thesis, organized around the five research hypotheses formulated in Chapter 2. It then discusses the broader theoretical and methodological implications and identifies directions for future research.

7.1 Main Findings

The following subsections assess each hypothesis in light of the empirical evidence gathered across the five publications.

H1: Clustering algorithms can uncover meaningful structure in mental health-related social media data – Partially confirmed

The clustering comparison in Publication I (Chapter 4.1) demonstrated that unsupervised algorithms can identify thematic structure in mental health-related Reddit posts: the best configuration (GMM with RoBERTa embeddings) discovered clusters corresponding to relationships, video games, and substance-related discussions, with the relationship cluster showing a higher proportion of depressed users. However, clustering did not separate diagnostic categories effectively (ARI = 0.180, NMI = 0.200), confirming that depression does not manifest as a geometrically separable class in any of the embedding spaces tested.

The most consequential finding was methodological: internal validation metrics (Silhouette Coefficient, Calinski-Harabasz, Davies-Bouldin) are not comparable across representation spaces with different geometric properties. In our experiments, TF vectors –whose magni-

tude grows with document length– consistently maximized internal metrics while performing worst on external validation. Naïve reliance on internal metrics would have selected the worst representation method, illustrating a failure mode that is likely to affect any study comparing clusterings across heterogeneous embedding spaces.

H2: Neural topic models produce more coherent topics than classical models on social media text – Confirmed

Publication II (Chapter 4.2) confirmed that BERTopic significantly outperforms LDA and TopClus in human evaluation of topic coherence on mental health social media data (average ratings: 2.49 vs. 1.14 vs. 1.61 on a 3-point scale, with inter-annotator agreement $\kappa \approx 0.50$). BERTopic's advantage stems from its use of pre-trained sentence transformers, which handle the lexical variability of informal text more effectively than LDA's bag-of-words representations. It should be noted that LDA was evaluated with default hyperparameters, a caveat that may underestimate its potential under careful tuning.

The study also confirmed that automatic coherence metrics provide an incomplete picture of topic quality. On the eRisk mental health data, NPMI achieved a Spearman correlation of only 0.112 with human judgments, while UMass (0.001), UCI (0.04), and C_V (0.05) were essentially uncorrelated. This failure is domain-specific: the same metrics achieve reasonable correlations on standard benchmarks (NPMI with Wikipedia: 0.735, as demonstrated in Publication III). The discrepancy arises because these metrics depend on co-occurrence statistics from a reference corpus, and the informal, emotionally charged vocabulary of mental health social media diverges substantially from the text these metrics were designed for. Human evaluation remains necessary for assessing topic quality in this domain.

Beyond model comparison, the discovered topics proved clinically relevant: BDI-II analysis revealed strong associations between specific topics (self-harm, relationship difficulties, weight loss) and depression indicators, and a temporal analysis captured the increase in depression-related discourse during the 2020–2021 pandemic period.

H3: Word embedding-based metrics offer a competitive alternative to classical metrics for topic coherence – Confirmed

Publication III (Chapter 5.1) demonstrated that GloVe embeddings (with dot product similarity) achieve an average Spearman correlation of 0.797 with human coherence judgments,

surpassing the best classical metric (NPMI with Wikipedia: 0.735). GloVe's performance was consistent across three diverse benchmarks (20 Newsgroups, New York Times, Genomics), with correlations exceeding 0.74 in every case –a level of cross-domain robustness that no classical metric matched. TBuckets, a method based on SVD decomposition and integer linear programming, achieved slightly higher correlation (0.806), offering a complementary algebraic perspective.

Contextual embeddings performed poorly for this task. BERT achieved a negative average correlation (-0.131), because its architecture requires sentence-level context to produce meaningful representations; when words are presented in isolation –as in coherence evaluation– the model lacks the input it was designed to exploit. This finding establishes a clear practical guideline: for word-level semantic tasks, static embeddings remain preferable to contextual models.

An important caveat applies: these results were obtained on standard benchmarks of well-edited text. The mental health social media domain that motivated the research was not included in the evaluation. Whether GloVe-based coherence maintains its advantage on noisy, informal mental health text remains an open question, though the cross-domain robustness results are encouraging.

H4: Temporal word embeddings can capture linguistic changes indicative of psychological disorders – Partially confirmed

Publication IV (Chapter 4.3) confirmed that temporal information carries genuine diagnostic value: the temporal embedding approach achieved competitive results across 18 eRisk datasets and four disorders, excelling on gambling detection (F1 = 0.84 with DCWE-SVM) and producing respectable results on anorexia and self-harm. Keyword analysis revealed that positive users exhibit more pronounced and irregular word movements than controls, and that the temporal signatures are disorder-specific.

However, the approach has a fundamental limitation. The temporal embedding methods (TWEC, DCWE) were designed for diachronic semantic change –tracking how word meanings evolve over years in society– and are repurposed here for a task where the relevant signal is emotional-contextual change in individual users. A depressed user does not use "sleep" in a new sense; they use it in a distressed context. The framework captures this shift only indirectly, which helps explain why it works well for gambling (where domain-specific vocabulary is

introduced) but poorly for depression (F1 = 0.35), where the change lies in the emotional tone of common words rather than in the introduction of new vocabulary.

Additional limitations include the dense-to-sparse information bottleneck (expensive dense embeddings are reduced to scalar delta features, discarding directional information) and vocabulary sensitivity (features are defined by training vocabulary, so test users whose language falls outside it are systematically disadvantaged). These limitations do not negate the core finding –temporal linguistic dynamics carry diagnostic signal– but they indicate that tools specifically designed for emotional-contextual tracking would be more effective than repurposed diachronic methods.

H5: A modular pipeline framework with automatic caching improves reproducibility and efficiency – Confirmed

Publication V (Chapter 4.4) demonstrated that LabChain's pipes-and-filters architecture with content-addressable caching delivers concrete improvements in research practice. The modular re-implementation of the temporal embedding pipeline from Publication IV revealed a preprocessing bug (failure to flatten hierarchical data) that had remained hidden in monolithic code; fixing it, combined with systematic exploration enabled by caching, produced performance improvements of up to 192.3% on some tasks. Caching eliminated approximately 48 hours of redundant computation across the mental health detection experiments.

LabChain's design –JSON-serializable pipelines as executable specifications, deterministic hash-based caching, and separation of public configuration from private state– makes reproducible, efficient experimentation the path of least resistance. The framework is published as open-source software and used in production research workflows at CiTIUS.

Transversal finding

Across all five publications, a recurring principle emerged: **evaluation is not separable from representation**. Internal clustering metrics depend on the geometry of the embedding space. Coherence metrics depend on the reference corpus. Temporal features depend on the training vocabulary. In every case, the metric's reliability is contingent on properties of the representational context that are often invisible to the practitioner. Recognizing and accounting for these dependencies is essential for drawing valid conclusions from computational text analysis.

7.2 Theoretical and Methodological Implications

The findings of this thesis carry implications across several areas of research.

For Information Retrieval and Text Mining. The demonstrated geometry-dependence of internal clustering metrics (Publication I) implies that comparative clustering studies across different representation spaces require external validation; internal metrics alone can be actively misleading. The embedding-based coherence metrics evaluated in Publication III provide a practical alternative to classical co-occurrence metrics: GloVe-based coherence is precomputed, requires no reference corpus at evaluation time, and shows superior cross-domain robustness. Researchers evaluating topic models should consider it as a default metric, while remaining aware that it shifts the dependency from the reference corpus to the pre-training data of the embedding model.

For mental health analysis from social media. BERTopic is more suitable than LDA for topic discovery on informal social media text, primarily due to its robustness to lexical variability (Publication II). However, the detectability of a mental health condition depends partly on its linguistic distinctiveness: conditions with focused, domain-specific vocabulary (gambling) are more amenable to word-level detection than conditions with diffuse linguistic signatures (depression) (Publication IV). This observation suggests that depression detection may require methods operating at the sentence or document level rather than the word level. Furthermore, the temporal embedding results indicate that while temporal linguistic dynamics carry diagnostic signal, the most productive temporal dimension for mental health is emotional-contextual change, not diachronic semantic change. Future temporal approaches should be designed around affect tracking rather than meaning evolution.

For topic model evaluation. Human evaluation remains necessary for assessing topic quality in specialized domains. The failure of automatic coherence metrics on mental health social media data (Publication II), contrasted with their reasonable performance on standard benchmarks (Publication III), demonstrates that the reliability of these metrics is domain-dependent. This domain-dependence is not a flaw of the metrics per se, but a consequence of their reliance on co-occurrence statistics that degrade when the reference corpus diverges from the target domain. Embedding-based metrics reduce but do not eliminate this dependency.

For reproducible research. LabChain's impact on this thesis –bug discovery, performance improvement, computational savings– illustrates that software engineering practices (modularization, caching, serialization) are not mere conveniences but can directly amplify scientific quality (Publication V). The finding that modular re-implementation exposed a hid-

den preprocessing error is a concrete argument for treating experimental code with the same engineering rigor applied to production software.

7.3 Future Research Directions

Several concrete research directions emerge from the findings and limitations of this thesis.

Coherence benchmarks for mental health text. The most immediate gap is the absence of human-annotated topic coherence benchmarks on mental health social media data. Constructing such benchmarks –with annotations from both NLP researchers and clinical psychologists– would allow direct evaluation of embedding-based coherence metrics in the domain that motivated their development, closing the argumentative gap between Publications II and III.

Dense temporal architectures with affect modeling. The limitations of the diachronic approach for mental health detection (Publication IV) point toward architectures that track emotional trajectories directly on dense representations. Transformer-based models with temporal attention could process sequences of sentence or document embeddings over time, preserving directional information and eliminating the vocabulary dependency of the current delta-based approach. The key shift is from asking "how has this word's meaning changed?" to "how has this user's emotional tone evolved?"

Fair comparison of LDA and BERTopic. The topic modeling comparison in Publication II used default LDA hyperparameters. A systematic comparison with properly tuned LDA –including hyperparameter optimization for number of topics, Dirichlet priors, and preprocessing– would establish more definitively the extent of BERTopic's advantage on social media text.

Clinical evaluation. The clinical relevance of automatically discovered topics was assessed via LLM-based BDI-II scoring (Publication II). Involving clinical psychologists in the evaluation of topics and temporal trajectories would strengthen the bridge between computational analysis and clinical utility, and could reveal aspects of topic quality that neither automatic metrics nor non-expert annotators capture.

Multilingual and multi-platform extension. All mental health experiments in this thesis used English-language Reddit data from eRisk. Extending the methods to other languages, platforms (Twitter/X, clinical forums), and cultural contexts is necessary to assess generalizability. This is particularly relevant for the temporal embedding approach, where language-specific

and platform-specific patterns may substantially affect performance.

LabChain development. Two directions merit attention: incorporating source code versioning into cache keys (to automatically invalidate cached results when library behavior changes) and enabling remote injection and distributed execution of pipelines. Because LabChain's content-addressable caching already decouples computation from storage, a natural extension is to allow pipelines to be serialized, transmitted, and executed on remote machines without requiring the full codebase to be present –leveraging the shared cache storage to coordinate inputs and outputs across distributed environments.

Bibliography

[1] Abdulrahman Aldkheel and Lina Zhou. Depression detection on social media: A classification framework and research challenges and opportunities. *Journal of Healthcare Informatics Research*, 8(1):88–120, 2024.

[2] Rubayyi Alghamdi and Khalid Alfalqi. A survey of topic modeling in text mining. *International Journal of Advanced Computer Science and Applications*, 6(1), 2015.

[3] Moez Ali. PyCaret: An open source, low-code machine learning library in Python. rlhttps://www.pycaret.org, 2020. PyCaret version 1.0.0.

[4] Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: Ordering points to identify the clustering structure. *ACM Sigmod record*, 28(2):49–60, 1999.

[5] Mario Ezra Aragon, Adrián Pastor Lopez-Monroy, Luis-Carlos Gonzalez Gonzalez-Gurrola, and Manuel Montes. Detecting mental disorders in social media through emotional patterns-the case of anorexia and depression. *IEEE Transactions on Affective Computing*, 2021.

[6] Robert Baggen, José Pedro Correia, Katrin Schill, and Joost Visser. Standardized Code Quality Benchmarking for Improving Software Maintainability. *Software Quality Journal*, 20(2):287–307, 2012.

[7] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc., 2009.

[8] David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[9] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[10] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information, June 2017. arXiv:1607.04606 [cs].

[11] Jordan Boyd-Graber, Yuening Hu, David Mimno, et al. Applications of topic models. *Foundations and Trends in Information Retrieval*, 11(2-3):143–296, 2017.

[12] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

[13] Manuel Couto and Anxo Perez and Javier Parapar and David E. Losada. Temporal Word Embeddings for Early Detection of Psychological Disorders on Social Media. *Journal of Healthcare Informatics Research*, pages 1–30, 2025.

[14] Jenny Chim, Adam Tsakalidis, Dimitris Gkoumas, Dana Atzil-Slonim, Yaakov Ophir, Ayah Zirikly, Philip Resnik, and Maria Liakata. Overview of the CLPsych 2024 shared task: Leveraging large language models to identify evidence of suicidality risk in online posts. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 177–190, 2024.

[15] Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. CLPsych 2015 shared task: Depression and PTSD on twitter. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 31–39, 2015.

[16] Manuel Couto, Javier Parapar, and David E. Losada. Comparison of clustering algorithms for knowledge discovery in social media publications: A case study of mental health analysis. *Procesamiento del lenguaje natural*, 73:69–81, 2024.

[17] Manuel Couto, Javier Parapar, and David E. Losada. Exploiting topic analysis models to explore psychological dimensions in social media data. *Scientific Reports*, 2025.

[18] Manuel Couto, Javier Parapar, and David E. Losada. LabChain: Enabling reproducible and modular scientific experiments in Python. *SoftwareX*, 2025.

[19] Manuel Couto, Javier Parapar, and David E. Losada. A study of word embedding models for measuring topic coherence. *Knowledge and Information Systems (under review)*, 2025.

[20] Fabio Crestani, David E Losada, and Javier Parapar. *Early Detection of Mental Health Disorders by Social Media Monitoring: The First Five Years of the eRisk Project*, volume 1018. Springer Nature, 2022.

[21] David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 1979.

[22] William HE Day and Herbert Edelsbrunner. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification*, 1(1):7–24, 1984.

[23] Munmun De Choudhury, Scott Counts, and Eric Horvitz. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th annual ACM web science conference*, pages 47–56, 2013.

[24] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

[25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[26] Valerio Di Carlo, Federico Bianchi, and Matteo Palmonari. Training temporal word embeddings with a compass. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6326–6334, 2019.

[27] Caitlin Doogan and Wray Buntine. Topic model or topic twaddle? re-evaluating demantic interpretability measures. In *North American Association for Computational Linguistics 2021*, pages 3824–3848. Association for Computational Linguistics (ACL), 2021.

[28] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

[29] Absalom E Ezugwu, Abiodun M Ikotun, Olaide O Oyelade, Laith Abualigah, Jeffery O Agushaka, Christopher I Eke, and Andronicus A Akinyelu. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, 110:104743, 2022.

[30] Adil Fahad, Najlaa Alshatri, Zahir Tari, Abdullah Alamri, Ibrahim Khalil, Albert Y Zomaya, Sebti Foufou, and Abdelaziz Bouras. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE transactions on emerging topics in computing*, 2(3):267–279, 2014.

[31] Maarten Grootendorst. BERTopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[32] Aron Halfin. Depression: the benefits of early and appropriate treatment. *American Journal of Managed Care*, 13(4):S92, 2007.

[33] William L Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*, 2016.

[34] Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. Dynamic contextualized word embeddings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2021.

[35] Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. Is automated topic model evaluation broken? the incoherence of coherence. *Advances in neural information processing systems*, 34:2018–2033, 2021.

[36] David E Losada, Fabio Crestani, and Javier Parapar. erisk 2017: Clef lab on early risk prediction on the internet: experimental foundations. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11–14, 2017, Proceedings 8*, pages 346–360. Springer, 2017.

[37] David E Losada, Fabio Crestani, and Javier Parapar. Overview of erisk: early risk prediction on the internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018, Proceedings 9*, pages 343–361. Springer, 2018.

[38] Veronica Lynn, Alissa Goodman, Kate Niederhoffer, Kate Loveys, Philip Resnik, and H Andrew Schwartz. CLPsych 2018 shared task: Predicting current and future psychological health from childhood essays. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 37–46, 2018.

[39] J MacQueen. Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pages 281–297. University of California Los Angeles LA USA, 1967.

[40] Aaron J Masino, Daniel Forsyth, and Alexander G Fiks. Detecting adverse drug reactions on twitter with convolutional neural networks and word embedding features. *Journal of Healthcare Informatics Research*, 2:25–43, 2018.

[41] Yu Meng, Yunyi Zhang, Jiaxin Huang, Yu Zhang, and Jiawei Han. Topic discovery via latent space clustering of pretrained language model representations. In *Proceedings of the ACM Web Conference 2022*, pages 3143–3152, 2022.

[42] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space, September 2013. arXiv:1301.3781 [cs].

[43] David N Milne, Glen Pink, Ben Hachey, and Rafael A Calvo. CLPsych 2016 shared task: Triaging content in online peer-support forums. In *Proceedings of the third workshop on computational linguistics and clinical psychology*, pages 118–127, 2016.

[44] Felix Mölder, Kim Philipp Jablonski, Brice Letcher, Michael B Hall, Christopher H Tomkins-Tinch, Vanessa Sochat, Jan Forster, Soohyun Lee, Sven O Twardziok, Alexander Kanitz, et al. Sustainable Data Analysis with Snakemake. *F1000Research*, 10:33, 2021.

[45] Thin Nguyen, Bridianne O'Dea, Mark Larsen, Dinh Phung, Svetha Venkatesh, and Helen Christensen. Using linguistic and topic analysis to classify sub-groups of online depression communities. *Multimedia tools and applications*, 76:10653–10676, 2017.

[46] Thin Nguyen, Dinh Phung, Bo Dao, Svetha Venkatesh, and Michael Berk. Affective and content analysis of online depression communities. *IEEE transactions on affective computing*, 5(3):217–226, 2014.

[47] Javier Parapar, Patricia Martín-Rodilla, David E. Losada, and Fabio Crestani. eRisk 2022: pathological gambling, depression, and eating disorder challenges. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part II*, pages 436–442. Springer, 2022.

[48] Javier Parapar, Patricia Martín-Rodilla, David E. Losada, and Fabio Crestani. Overview of eRisk 2023: Early risk prediction on the internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 294–315. Springer, 2023.

[49] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[50] Michael J Paul and Mark Dredze. Discovering health topics in social media using topic models. *PloS one*, 9(8):e103408, 2014.

[51] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine Learning in Python. *the Journal of Machine Learning Research*, 12:2825–2830, 2011.

[52] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of LIWC2015. Technical report, 2015.

[53] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global Vectors for Word Representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[54] A Picardi, I Lega, L Tarsitani, M Caredda, G Matteucci, MP Zerella, R Miglio, A Gigantesco, M Cerbo, A Gaddini, et al. A randomised controlled trial of the effectiveness of a program for early detection and treatment of depression in primary care. *Journal of affective disorders*, 198:96–101, 2016.

[55] QuantumBlack, a McKinsey company. Kedro: An Open-source Python Framework for Creating Reproducible, Maintainable and Modular Data Science Code. `https://kedro.org`, 2019. Accessed: 2025-05-23.

[56] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[57] Radim Řehůřek and Petr Sojka. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50. ELRA, 2010.

[58] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[59] Douglas A Reynolds. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663), 2009.

[60] Marcel Rieger, Martin Erdmann, Benjamin Fischer, and Robert Fischer. Design and Execution of Make-like, Distributed Analyses Based on Spotify's Pipelining Package Luigi. *arXiv preprint arXiv:1706.00955*, 2017.

[61] Esteban A. Ríssola, David E. Losada, and Fabio Crestani. A survey of computational methods for online mental state assessment on social media. *ACM Trans. Comput. Healthcare*, 2(2), mar 2021.

[62] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[63] Maja Rudolph and David Blei. Dynamic embeddings for language evolution. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 1003–1011, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee.

[64] Ariel Shensa, Jaime E Sidani, Mary Amanda Dew, César G Escobar-Viera, and Brian A Primack. Social media use and depression and anxiety symptoms: A cluster analysis. *American journal of health behavior*, 42(2):116–128, 2018.

[65] Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617, 2002.

[66] World Health Organization. Depressive disorder (depression), 2023. `https://www.who.int/news-room/fact-sheets/detail/depression`, Last accessed on 2024-01-09.

[67] Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. Dynamic word embeddings for evolving semantic discovery. pages 673–681, 02 2018.

[68] Amir Hossein Yazdavar, Hussein S Al-Olimat, Monireh Ebrahimi, Goonmeet Bajaj, Tanvi Banerjee, Krishnaprasad Thirunarayan, Jyotishman Pathak, and Amit Sheth. Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*, pages 1191–1198, 2017.

[69] Ayah Zirikly, Philip Resnik, Ozlem Uzuner, and Kristy Hollingshead. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the sixth workshop on computational linguistics and clinical psychology*, pages 24–33, 2019.

List of Figures

List of Tables

Mental health disorders are a growing global concern, and online platforms have become spaces where individuals leave linguistic traces of psychological distress. The vast amount of social media text provides an opportunity to study these conditions at scale. Unsupervised techniques enable the discovery of latent structures associated with specific disorders, offering insights into how language reflects mental health. This knowledge can be leveraged through predictive models to identify at-risk individuals. However, the reliability of these approaches depends critically on evaluation, as metrics and representations are deeply intertwined. In this thesis, we present advances across the pipeline, from text representation and unsupervised analysis to temporal modeling, evaluation, and reproducible research.