Knowledge and Information Systems (2026) 68:160
https://doi.org/10.1007/s10115-026-02782-6

RESEARCH

A study of word embedding models for measuring topic coherence

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

Received: 6 May 2025 / Revised: 16 February 2026 / Accepted: 23 April 2026

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

✉ David E. Losada
david.losada@usc.es

Manuel Couto
manuel.couto.pintos@usc.es

Javier Parapar
javier.parapar@udc.es

[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

Published online: 19 May 2026

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.

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 Fig. 1).

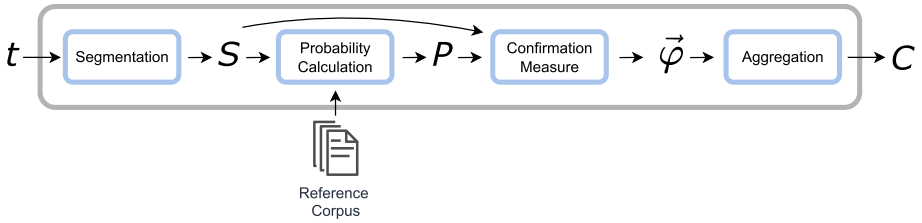


Fig. 1 Diagram illustrating the steps to compute the coherence (C) of a topic (t) [31]

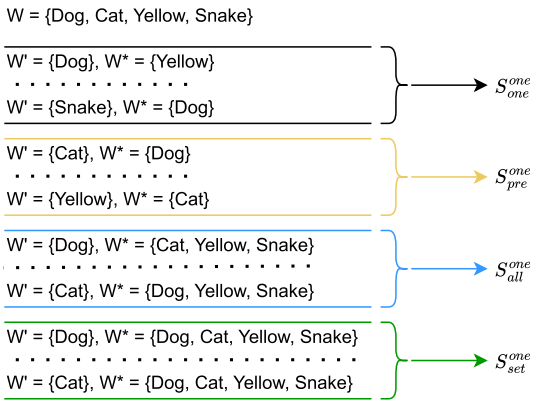


Fig. 2 Examples of segmentation techniques applied to a topic whose top words are "Dog", "Cat", "Yellow", and "Snake"

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.

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:

$$S^{one}_{all}(W) = \left\{(W', W^*) | W' = \{w_i\};\, w_i \in W;\, W^* = W \setminus \{w_i\}\right\} \tag{1}$$

$$S^{one}_{set}(W) = \left\{(W', W^*) | W' = \{w_i\};\, w_i \in W;\, W^* = W\right\} \tag{2}$$

$$S^{one}_{one}(W) = \left\{(W', W^*)|W' = \{w_i\};\ W^* = \{w_j\};\ w_i, w_j \in W; i \neq j\right\} \quad (3)$$

$$S^{one}_{pre}(W) = \left\{(W', W^*)|W' = \{w_i\};\ W^* = \{w_j\};\ w_i, w_j \in W; i > j\right\} \quad (4)$$

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

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

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

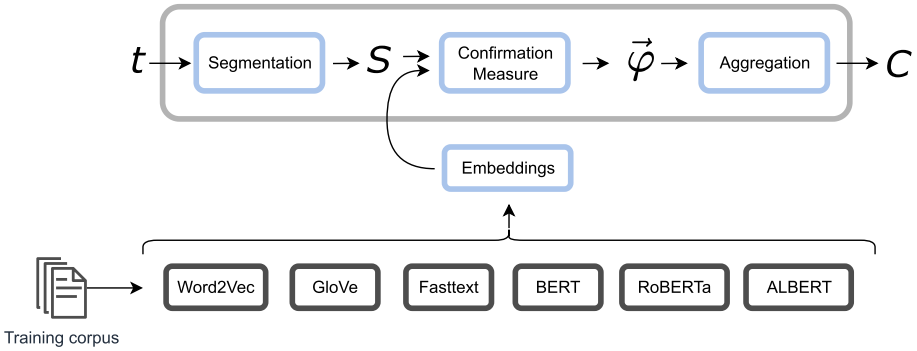


Fig. 3 Diagram illustrating the computation of coherence (C) for a topic (t) based on embeddings

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 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 nonlinearity 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 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 pretraining. 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,

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 ‡

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	**20NG**	**NYT**	**Geno**	**Avg.**
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**

(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

[1] The results of Tbuckets, due to the difficulties in reproducing this method, have been taken directly from [15].

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.

- **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, Fig. 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,

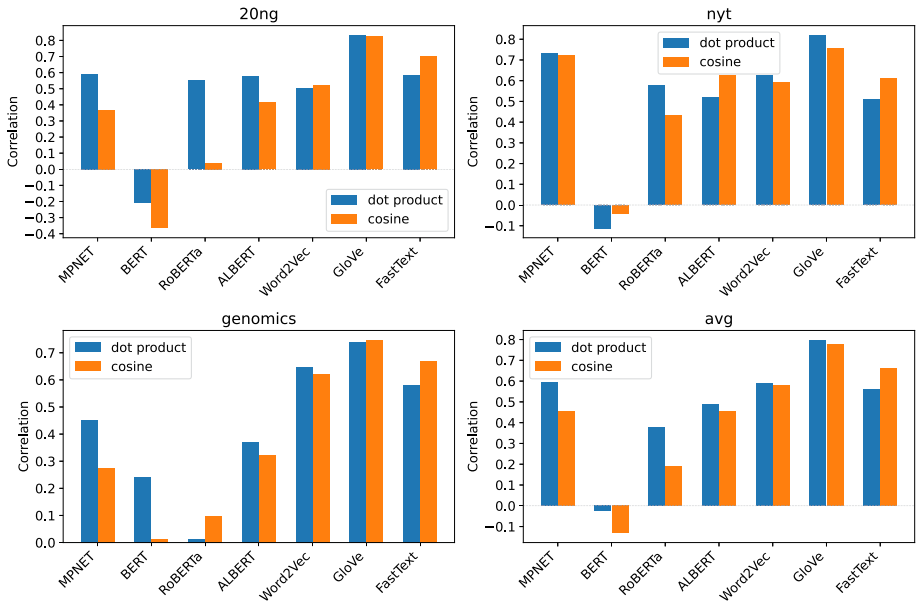


Fig. 4 Comparison of embedding models with different similarity metrics (cosine and dot product)

while other models (e.g., BERT) seem to be better suited for dot product. The best performing model, GloVe, performs well with both alternatives.

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 nonlinear 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 which performance steadily degrades. This pattern suggests that MPNet's early layers effectively capture the semantic

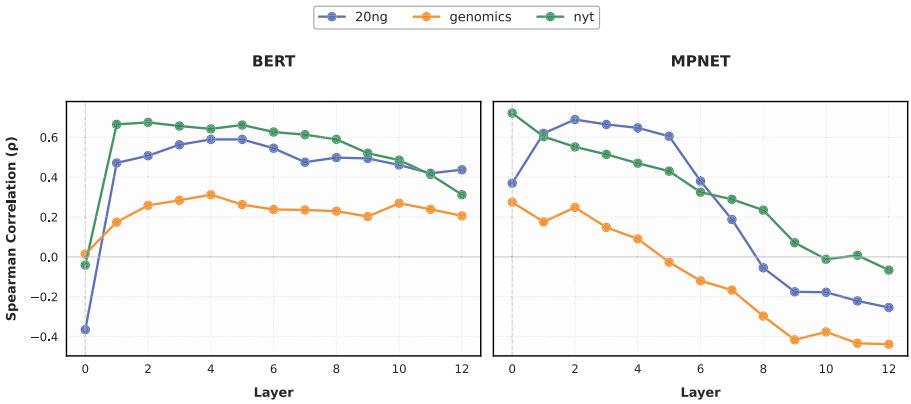


Fig. 5 Extracting word embeddings from different transformer layers. Effect on topic coherence performance

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 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 (Fig. 6), the x-axis corresponds to human rankings and the y-axis corresponds to model rankings. Topics near the diagonal indicate accurate relative

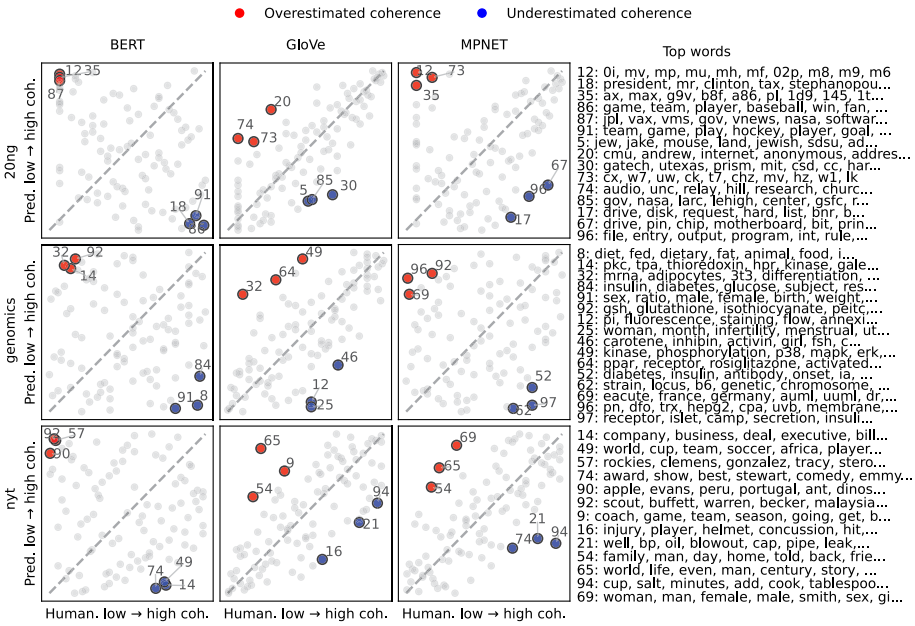


Fig. 6 The blue (red) topics represent topics in which human evaluators perceived high (low) coherence, but the models estimated low (high) coherence

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.

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 Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. 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 Availability 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 Statement The code to reproduce the experiments is available at Github.

Declarations

Conflict of interest The authors declare no conflict of interest.

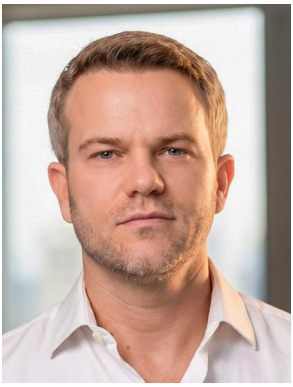
References

1. Blei DM (2012) Probabilistic topic models. Commun ACM 55(4):77–84
2. Boyd-Graber J, Hu Y, Mimno D et al (2017) Applications of topic models. Found Trends® Inf Retrieval 11(2–3):143–296
3. Alghamdi R, Alfalqi K (2015) A survey of topic modeling in text mining. Int J Adv Comput Sci Appl (IJACSA) 6(1)
4. Xu W, Hirami K, Eguchi K (2025) Self-supervised learning for neural topic models with variance–invariance–covariance regularization. Knowledge Inf Syst, 1–19
5. Hoyle A, Goel P, Hian-Cheong A, Peskov D, Boyd-Graber J, Resnik P (2021) Is automated topic model evaluation broken? the incoherence of coherence. Adv Neural Inf Process Syst 34:2018–2033

6. Rosner F, Hinneburg A, Röder M, Nettling M, Both A (2014) Evaluating topic coherence measures. arXiv. arXiv:1403.6397
7. Rahimi H, Hoover JL, Mimno D, Naacke H, Constantin C, Amann B (2023)Contextualized topic coherence metrics. arXiv preprint arXiv:2305.14587
8. Zhou R, Wang F, Liu C, Lu X (2025) Sparse dynamic topic model with topic birth and death over time. Knowledge Inform Syst, 1–23
9. Chang C-H, Hwang S-Y (2021) A word embedding-based approach to cross-lingual topic modeling. Knowl Inf Syst 63(6):1529–1555
10. Gao W, Peng M, Wang H, Zhang Y, Xie Q, Tian G (2019) Incorporating word embeddings into topic modeling of short text. Knowl Inf Syst 61:1123–1145
11. Du C, Yao K, Zhu H, Wang D, Zhuang F, Xiong H (2024) Mining technology trends in scientific publications: a graph propagated neural topic modeling approach. Knowl Inf Syst 66(5):3085–3114
12. Nikolenko SI (2016) 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
13. Fang A, Macdonald C, Ounis I, Habel P (2016) 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
14. Belford M, Greene D (2019) Comparison of embedding techniques for topic modeling coherence measures. In: LDK (Posters), pp. 1–5
15. Ramrakhiyani N, Pawar S, Hingmire S, Palshikar G (2017) 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
16. Chang J, Gerrish S, Wang C, Boyd-Graber J, Blei D (2009) Reading tea leaves: How humans interpret topic models. Adv Neural Inf Process Syst 22
17. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022
18. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R (1990) Indexing by latent semantic analysis. J Am Soc Inf Sci 41(6):391–407
19. AlSumait L, Barbará, D, Gentle J, Domeniconi C (2009) 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. Springer
20. Newman D, Lau JH, Grieser K, Baldwin T (2010) 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
21. Mimno D, Wallach H, Talley E, Leenders M, McCallum A (2011) Optimizing semantic coherence in topic models. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 262–272
22. Aletras N, Stevenson M (2013) 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. https://aclanthology.org/W13-0102
23. Röder M, Both A, Hinneburg A (2015) Exploring the space of topic coherence measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, pp. 399–408
24. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient Estimation of Word Representations in Vector Space. arXiv. arXiv:1301.3781 [cs]
25. Pennington J, Socher R, Manning C (2014) 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), Association for Computational Linguistics, Doha, Qatar, pp 1532–1543
26. Bojanowski P, Grave E, Joulin A, Mikolov T (2017)Enriching Word Vectors with Subword Information. arXiv. arXiv:1607.04606 [cs]
27. Doogan C, Buntine W (2021) Topic model or topic twaddle? re-evaluating demantic interpretability measures. In: North American Association for Computational Linguistics 2021, pp. 3824–3848. Association for Computational Linguistics (ACL)
28. Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: Bidirectional encoder representations from transformers. arXiv preprint arXiv:1810.04805, 15
29. Liu Y (2019) Roberta: A robustly optimized bert pretraining approach. arXiv preprint **364**arxiv:1907.11692
30. Lan, Z (2019) Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942

31. Röder M, Both A, Hinneburg A (2015) 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
32. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser LU, Polosukhin I (2017) Attention is all you need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in Neural Information Processing Systems, vol 30. Curran Associates Inc, USA
33. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv. arXiv:1907.11692 [cs]
34. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R (2020) ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. arXiv. arXiv:1909.11942 [cs]
35. Song K, Tan X, Qin T, Lu J, Liu T-Y (2020) MPNET: masked and permuted pre-training for language understanding. Adv Neural Inf Process Syst 33:16857–16867
36. Bengio Y, Ducharme R, Vincent P, Jauvin C (2003) A neural probabilistic language model. J Mach Learn Res 3:1137–1155
37. Hinton GE (1984) Distributed representations
38. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323(6088):533–536
39. Elman JL (1990) Finding structure in time. Cogn Sci 14(2):179–211
40. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV (2019) Xlnet: Generalized autoregressive pretraining for language understanding. Adv Neural Inf Proc Syst 32
41. Bianchi F, Terragni S, Hovy D, Nozza D, Fersini E (2021) 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
42. Grootendorst M (2022) BERTopic: Neural topic modeling with a class-based tf-idf procedure. arXiv preprint arXiv:2203.05794
43. 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) Association for Computational Linguistics, Hong Kong, China, pp 3982–3992
44. Jawahar G, Sagot B, Seddah D (2019) 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
45. Tenney I, Das D, Pavlick E (2019) 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

Manuel Couto is a PhD candidate at the Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS) of the Universidad de Santiago de Compostela (Spain). He received his BSc in Computer Science in 2021 and his MSc in Computer Science in 2023 from the Facultad de Informática (FIC) at the Universidad de A Coruña (UDC). He spent one year as a research assistant at the Laboratorio de Recuperación de Información (IRLab, UDC) before joining CiTIUS in 2023, where he is currently pursuing his PhD under the supervision of David E. Losada and Javier Parapar. His research interests lie at the intersection of Natural Language Processing, Machine Learning, and computational mental health analysis, with particular emphasis on early risk detection on social media, temporal word embeddings, topic modeling for psychological dimensions, and reproducible scientific experimentation.



Javier Parapar is an Associate Professor of Computer Science and Artificial Intelligence at the University of A Coruña (Spain), where he leads the Information Retrieval Lab and is affiliated with CITIC. His research interests cover Information Retrieval, Natural Language Processing, and Artificial Intelligence, with a particular focus on socially relevant and health-related applications such as early detection of mental health risks, hate speech detection, health misinformation detection, and the evaluation of language models. His work places special emphasis on explainability, reliability, and robust evaluation methodologies for AI and IR systems. He regularly publishes in leading international venues in the field, including ACM SIGIR, ECIR, ACM RecSys, and ACL conferences, and is actively involved in European research initiatives, including Marie Skłodowska-Curie Actions. He combines academic research with knowledge transfer activities and the organization of international scientific events.

David E. Losada is a Full Professor of Computer Science and Artificial Intelligence at the University of Santiago de Compostela (Spain). He received his BS in Computer Science (with honors) in 1997 and his PhD (with honors) in 2001 from the University of A Coruña. He was a lecturer at San Pablo-CEU University (2001–2002) and joined the University of Santiago de ompostela in 2003 as a Ramón y Cajal Senior Research Fellow. His research interests cover Information Retrieval and related areas, including IR evaluation, early risk detection, misinformation detection, IR models, summarization, novelty detection, sentence retrieval, and opinion mining. He is an active member of the IR community, regularly serving on the Program Committees of leading international conferences such as SIGIR and ECIR, has led several R&D projects and contracts in search technologies, and was recognized as an ACM Senior Member in 2011.