Journal of Universal Computer Science, vol. 32, no. 11 (2026), 1560-1589
submitted: 25/5/2025, accepted: 21/2/2026, appeared: 28/10/2026 CC BY 4.0

Semantic Representations based on Neural Topic Models

Fagner Leal Pantoja
(Universidade Estadual de Campinas, Campinas, Brazil
https://orcid.org/0000-0003-1784-5512, fagnerleal@ufpa.br)

André Santanchè
(Universidade Estadual de Campinas, Campinas, Brazil
https://orcid.org/0000-0002-1756-4852, santanch@unicamp.br)

Claudia Bauzer Medeiros
(Universidade Estadual de Campinas, Campinas, Brazil
https://orcid.org/0000-0003-1908-4753, cmbm@unicamp.br)

Abstract: Topic Models is a tool to extract semantics from textual collections. However, in spite of extensive work on topic modeling, including its evaluation using Language Models (LM), there is need for more research to measure, analyze and validate the topics learned. To this end, we present a Neural Topic Model called ABT which leverages semantic vector spaces (derived from language models) to uncover latent semantic representations in text collections. ABT enables one to explore the themes covered by a given textual collection and analyze them as a hierarchical structure of topics. We demonstrate the effectiveness of ABT by analyzing two text corpora: (1) CliCR, a collection of 10,539 real clinical cases; and (2) MovieLens, a dataset consisting of 9,742 movie titles. We have experimented with different language models to investigate the effects of using specific tasks to tune the neural network. The topics that resulted from our study highlighted concepts that are shown to be effectively used as prominent topics to represent the datasets. The results show that each experimented LM focuses on distinct features to produce the topics and may generate topics with different grades of generality and specificity. Our study sheds light on recent findings of the related literature, which state that the topic coherence metric does not align with assessments performed by human evaluators.

Keywords: Natural Language Processing, Topic Modeling, Language Models, Transformers, Semantic Representation, Biomedical Data
Categories: H.3.1, H.3.3, I.2.4, I.2.7, I.5.3, J.3
DOI: 10.3897/jucs.160058

1 Introduction

Topic Modeling [Blei et al. 2023, Griffiths et al. 2007, Mimno et al. 2011, Hoyle et al. 2021, Grootendorst 2022] is a Natural Language Processing (NLP) task whose objective is to identify the main topics covered in a textual collection and encode them succinctly in a computational abstraction referred to as a Topic Model. A topic is said to be *coherent* when it brings together a group of words that make sense when used in the same context [Boyd-Graber et al. 2014], such as the Biomedical topic associated with the words: {myocardium, coronary, acute, artery, pain}.

A topic model provides an overview of a given textual collection, and is handy when massive amounts of text prevent a person from reading the entire collection. Furthermore,

the resulting topics may be helpful to specific problems in the NLP research agenda. For example, they can improve tasks such as Sentiment Analysis, Machine Translation, or Domain Adaptation of Language Models [Boyd-Graber et al. 2017].

Although there is extensive literature on topic models, there is still a lack of validation regarding the coherence, interpretability and organization of the produced topics. Therefore, in this work we analyze the hierarchical structure underlying the set of topics and the Topics Coherence measured in each level of the inferred structure.

To this effect, this article presents our Attention-based Topics (ABT) method, a Neural Topic Modeling approach [Angelov 2020, Grootendorst 2022] that identifies the major themes in a given textual collection. Our method produces topics based on Neural Language Models [Devlin et al. 2019, Reimers and Gurevych 2019] to represent a textual collection through multidimensional vectors.

ABT infers latent topics from contextual information extracted from a textual collection, combining two representations used in the process in a complementary way: the semantic vector space $\mathbb{R}^{768}$ (embedded on the neurons of the language model) and the matrix cTFIDF. These two representations jointly model statistical patterns observed in the underlying structure of the textual collection. Paraphrasing the concept of the *gist* of a document, proposed by Griffiths et al. [2007], we state that our approach encodes the gist of the input collection.

We have conducted two case studies of our approach, applied to the biomedical and movie domains, using respectively real data collections from the CliCR corpus [Šuster and Daelemans 2018] and the MovieLens [Harper and Konstan 2015] collection. We validated the quality of the produced topics by measuring the Topics Coherence metric [Röder et al. 2015], which assesses whether the most discriminant words in a set of topics reinforce each other.

Besides different kinds of quantitative evaluation against other methods (e.g., performance metrics, topic coherence metrics), we have qualitatively evaluated the behavior of ABT in terms of topic granularity. Our method produces topics organized in a Hierarchical Structure underlying the pre-trained language model. Our approach presents itself as an alternative view to the topic modeling problem, taking advantage of the context and linguistic knowledge internally incorporated into the neural network parameters to produce coherent topics with different granularity options along the path of the hierarchical structure.

The remainder of the text is organized as follows: section 2 discusses background foundations and related work; section 3 describes ABT; section 4 describes our test data and an overview of how they were processed; section 5 presents quantitative and qualitative results of a case study of our ABT approach tailored to the medical domain handling a collection of clinical cases. Finally, section 6 presents our concluding remarks.

To support reproducibility, appendix A presents the Python code of our method.

The complete source code to reproduce the reported experiments is available in the Zenodo repository at (https://doi.org/10.5281/zenodo.18745143).

The data underpinning the analysis reported in this paper are deposited at “Data repository” REDU at https://doi.org/10.25824/redu/R15PFJ.

2 Background and Literature Review

This section provides the basic framework for understanding our approach, providing references on basic definitions (subsection 2.1), a summary of neural topic modeling

(subsection 2.2), and some evaluation mechanisms for topic modeling (subsection 2.3). It concludes with a brief analysis of related work under this framework (subsection 2.4)

Additional related work is commented in sections 3 and 5, when results of other works are contrasted with ours.

2.1 Definitions and Terminology Adopted

Topic Modeling [Blei et al. 2023, Griffiths et al. 2007, Mimno et al. 2011, Hoyle et al. 2021, Grootendorst 2022, Abdelrazek et al. 2023] is an unsupervised Natural Language Processing challenge whose goal is to discover topics that summarize the textual collection under analysis. A topic model explicitly represents the latent semantic structure [Angelov 2020] – or gist [Griffiths et al. 2007] – of a textual collection.

A wide range of specific-domain applications can benefit from the topics discovered using such a model, such as health, cognition studies, social studies, bioinformatics, or education [Kennedy et al. 2021, Kulkarni et al. 2021, Hoyle et al. 2021, Boyd-Graber et al. 2017, Griffiths et al. 2007]. For instance, Kennedy et al. [2021] apply topic models to analyze the correlation between individual moral concerns and language usage. In another branch of research, Kulkarni et al. [2021] investigate the presence of Post Traumatic Stress Disorder in textual reports on Reddit[1] using topic models.

The concept of Topics Model refers to a discrete probability distribution describing the connections between words, topics, and documents [Boyd-Graber et al. 2017]. Topics are word combinations that demonstrate idiosyncrasies in the linguistic distribution of the corpus under analysis [Bouma 2009]. Topic models are explicit representations that probabilistically associate documents with topics and topics with words [Blei et al. 2023].

According to Abdelrazek et al. [2023], there are four categories of topic models: algebraic, fuzzy, probabilistic, and neural. These different categories offer different advantages and are suited for particular settings. Our work can be characterized as being part of neural topic models.

Many methodologies address the topic modeling problem, such as Latent Semantic Analysis (LSA) based on linear algebra and its probabilistic version pLSA. Such methods apply dimensionality reduction to the documents represented in a Bag-of-Words format. The bag-of-words representations are adequate since, by hypothesis, word order is not a determining factor in these methods [Blei et al. 2023].

Latent Dirichlet Allocation (LDA) [Blei et al. 2023] is a probabilistic model for discrete data collections, such as textual data. The LDA uses two complementary distributions: a topic over words distribution that describes the relationship between topics and words; and a second distribution that allocates topics to documents [Boyd-Graber et al. 2017]. For LDA, a document is a random mixture of latent topics, which in turn are probability distributions over vocabulary words. LDA is a dimensionality reduction technique that is considered one of the most robust and efficient methods for topic modeling, and applies to problems beyond the textual domain [Blei et al. 2023].

2.2 Neural Topic Modeling

Neural Topic Modeling (NTM) is a current research trend that combines topic modeling with Neural Language Models (e.g., Angelov [2020], Dieng et al. [2020], Thompson and Mimno [2020], Peinelt et al. [2020], Zhao et al. [2021], Hoyle et al. [2021], Asgari-Chenaghlu et al. [2021], Kulkarni et al. [2021], Grootendorst et al. [2022]). Indeed, this is

[1] https://www.reddit.com/

a rapidly growing area of research. Thompson and Mimno [2020] and Zhao et al. [2021] discuss the similarity between the topics produced via NTM and other traditional topic models, such as LDA. NTM-based works claim to produce more interpretable topics than prior methods, yielding improvements in the state-of-the-art concerning topic coherence measure [Hoyle et al. 2021].

Our proposed method, ABT, can be stated as a NTM approach. Thus, there follows a brief literature review of Language Models and Transformers-based neural networks, which form the basis of ABT.

2.2.1 Neural Language Models

A neural language model is a type of word inventory that represents a unit of textual information (such as a word, a sentence, a paragraph, or a document) through a numerical vector that encodes the most relevant characteristics of this textual unit that are typically induced by a neural network [Turian et al. 2010]. This set of relevant characteristics alludes to a type of semantics grounded on the Distributional Hypothesis [Martin and Jurafsky 2019], which deems that words that occur in the same context tend to have similar meanings.

Language models are also known as Large Language Models (LLM), Word Embeddings, Word Vectors, Feature Vectors or Word Representations [Bengio et al. 2000, Turian et al. 2010, Bojanowski et al. 2017].

One of the pioneer models of word representation is the One-Hot Vector, which represents each word as a vector of size $|W|$. Unfortunately, the one-hot representation is inefficient due to its high dimensionality (one dimension for each word in the vocabulary) and its nature as a sparse matrix model, as each vector is composed of 0 values in most of its dimensions [Turian et al. 2010].

Other representations were subseqently presented: Word2Vector [Mikolov et al. 2013a], GloVe [Pennington et al. 2014], ELMO [Peters et al. 2018], GPT [Radford et al. 2018], BERT [Devlin et al. 2019], Sentence-BERT [Reimers and Gurevych 2019], etc.

In pioneering work using neural networks to train word representations, Mikolov et al. [2013a], Mikolov et al. [2013b], and Mikolov et al. [2013c] presented neural networks to train language models to represent linguistic regularities in the form of a vector space. This vector space allows one to perform vector algebra operations and, therefore, infer a kind of semantic-statistical relationship between the words based on the displacements between the vectors of the textual collection modeled in the neural network. Another example of regularity captured through simple algebraic operations on the vectors of words is the similarity between the words big and bigger in the same way as between the words small and smaller [Mikolov et al. 2013a]. There are indications that the linguistic regularities captured by the neural models carry apparent syntactic and semantic aspects preserved in the statistical patterns encoded in the referred semantic vector space.

To the best of our knowledge, the Transformer is currently the most robust and suitable model for pre-training Language Models, which will be further explored in the next section. The Transformer is the core technology driving the NLP technological leap and, consequently, the so-called Artificial Intelligence (AI) revolution.

2.2.2 Transformers

The transformer model [Vaswani et al. 2017] is a deep neural network that efficiently solves long-date problems faced by the NLP research area. The Transformer architecture

implements the Attention Mechanism to infer the statistical context of a sentence as a whole – i.e., a sentence's global context. The attention mechanism solved a critical gap inherent to the previous models which has limited their contexts to a window of few terms around the focus.

The transformer produces a set of characteristic vectors, encoded in the matrix QKV [Vaswani et al. 2017] as a byproduct of the training activity [Luong et al. 2015]. While training, the characteristic vectors are updated with the contribution of all terms surrounding the current training term. In this way, the transformer is aware of the bi-directional context of each term at hand [Ruder 2019] and can guide the attention to the more discriminant terms in the entire training dataset [Luong et al. 2015] in both directions. Differently, the previously proposed approaches (e.g., LSTM) were limited to obtaining information just from one direction, left-to-right or right-to-left.

The characteristic vectors produced by the transformer can be reused as a universal representation to accomplish other tasks beyond the original task performed by the transformer, since the matrix QKV embeds syntactic and semantic aspects that are discriminant to many NLP tasks. During the training phase, the model perceives such aspects as idiosyncrasies and encodes them in the inferred vector. One can use this linguistic knowledge encapsulated in QKV to perform other tasks through Transfer Learning techniques [Howard and Ruder 2018].

The realization of such a universal representation capable of handling diverse tasks is another key aspect of the Transformer architecture, which ultimately helped drive the AI revolution.

2.2.3 BERT (Bidirectional Encoder Representations from Transformers)

BERT is a neural language model pre-trained on the Masked Language Modeling task through the Transformer architecture [Devlin et al. 2019]. The pre-training objective is to uncover the term hidden by a [MASK] symbol in the sentences given as input (such as "I like this [MASK] bed''). During this pre-training phase, the network embeds the bidirectional context associated with the statistical relationship between the given masked term with all its surrounding terms within the sentence. The context emerges from patterns learned in all the sample sentences previously observed by the model by performing the language modeling (during the pre-training phase). The language modeling captures a wide range of resources and facets – e.g., long-term dependencies, hierarchical relationships, sentiments, and patterns spiked at the writing – directly associated with the source task used in the pre-training [Howard and Ruder 2018].

BERT is a breakthrough model improving the state-of-the-art in many NLP tasks. To the best of our knowledge, BERT word vectors are one of the most powerful methods to transform textual information into an effective computational abstraction rich in context-sensitive information.

Internally, within the network parameters, BERT maintains a vector space $\mathbb{R}^H$ of multiple H dimensions. Usually, $H = 768$ depending on the selected version: BERT$_{base}$ (H=768, parameters=110M) or BERT$_{large}$ (H=1024, parameters=340M) [Devlin et al. 2019].

The BERT vector space $\mathbb{R}^H$ embeds a 30000 token vocabulary. To put it less abstractly, Figure 1 illustrates BERT space $\mathbb{R}^{768}$ in a 3-dimensional rendering obtained by using the Embedding Projector tool[2]. We have drawn this 3D picture after reducing

[2] https://projector.tensorflow.org/

the dimensions (through a PCA method) from $\mathbb{R}^{768}$ to $\mathbb{R}^3$, because it is impossible to draw/visualize the original 768 dimensions.

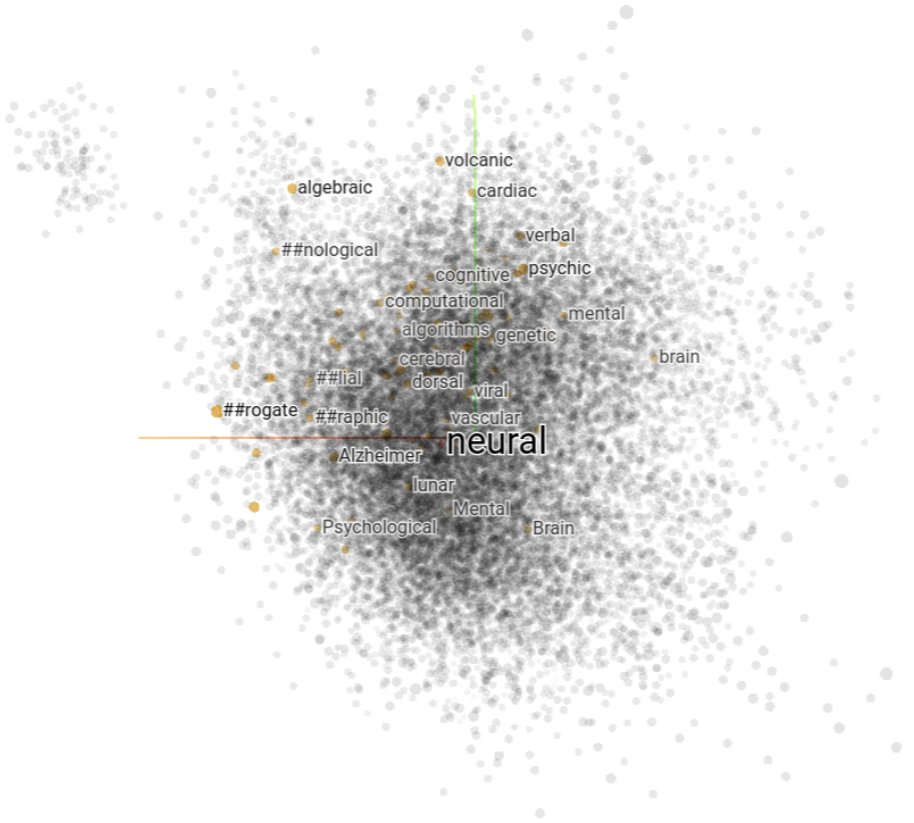

Figure 1: A 3-dimensional picture of the BERT vector space captured by reducing it from R^{768} dimensions into R^3. This space is internally written in the parameters of the network. In such a space, how close two terms are, more similar they are, and therefore probably occur in similar contexts.

In this vector space, the closer two terms are, the more similar they are, and therefore probably occur in similar contexts. The figure highlights the word "neural" and its more similar words. It is possible to note two distinguishable groups of words {cognitive, psychic, genetic, cerebral, Mental, …} and {computational, algorithms} standing for two main contexts: referring to the biological Neural System and to Artificial Neural Networks from the Computer Science terminology.

In Figure 1, it is also worth mentioning the words {##nological, ##rogate, ##raphic, ##lial} which are used to compose rare and out-of-vocabulary words. These terms are sub-word units [Wu 2016], a minimal set of terms that is capable of representing any arbitrary word. For example, the word Neurological is not in the vocabulary, then it will be processed into a sequence of sub-word units: {N, ##eur, ##ological}.

The development of such a minimal set of sub-words is another important aspect of the NLP technological leap, as creating a word inventory of this kind has long been a long-standing problem in NLP tasks.

2.2.4 Sentence-BERT

The arrival of word vector spaces (e.g., BERT) have inspired other approaches (e.g., SkipThought [Kiros et al. 2015], InferSent [Conneau et al. 2017], Sent2Vec [Pagliardini et al. 2017], and Sentence-BERT [Reimers and Gurevych 2019]) to generate language models for larger text segments, such as phrases, sentences, paragraphs, and even entire documents.

Sentence-BERT (SBERT) is a sentence language model [Reimers and Gurevych 2019] which uses siamese networks (i.e., networks with tied weights) to infer sentence vectors from the BERT word vectors. The training procedure of Sentence-BERT takes a pair of sentences as input, transforms each one into a sentence vector from the averaging of its word vectors, and computes the similarity between the two sentence vectors. In the end, both sentences' vectors are tuned according to this similarity function, so that similar pairs are closer and dissimilar pairs are further apart.

2.3 Some evaluation strategies for topic models

One important issue in our work is how we evaluate ABT. There are diverse approaches to assess the quality of topic models, such as human evaluation, automatic measures and more recently prompt-based evaluation. The evaluation approaches generally rely on quantitative and/or qualitative analysis. The metrics called Topics Coherence is the most adopted among quantitative strategies. For this reason, we have adopted it in our research.

Topics Coherence is a metric for automatically measuring the quality of a topic by measuring the interpretability of the topic model. The intuition is: a topic said to be coherent must contain words that are, in some way, associated [Boyd-Graber et al. 2014]. Usually, this association is measured based on the Pointwise Mutual Information (PMI) of two randomly sampled words from the same sentence [Dieng et al. 2020].

In our work, we have used the Gemsim framework[3] to measure topics' coherence [Röder et al. 2015]. We have used the c_v configuration which combines the indirect cosine measure with NPMI and Boolean Sliding Window. First, the top words W of a topic $t \in T$ are divided in the subsets W' and W^*, according to:

$$S_{set}^{one} = \{(W', W^*) | W' = w_i; w_i \in W; W^* = W\} \tag{1}$$

The subset W^* is used to confirm the subset W' by comparing words to the total word set W. Next, probabilities are estimated to W' and W^* based on word co-occurrence counts using the Boolean Sliding Window approach, in which a window moves over the input corpus C one word token per step. The j^{th} element of the context vector v_i of word w_i has NPMI:

$$v_{ij} = NPMI(w_i, w_j)^{\gamma} = \left(\frac{log \frac{P(w_i, w_j) + \epsilon}{P(w_i) \cdot (w_j)}}{-log(P(w_i, w_j) + \epsilon)} \right)^{\gamma} \tag{2}$$

[3] https://radimrehurek.com/gensim/models/coherencemodel.html

The indirect cosine is used to compute how strong the word set W^* supports W' comparing the similarity of the vectors $\vec{u}$ and $\vec{w}$ (computed with respect to the estimated probabilities), according to:

$$s_{cos}(\vec{u}, \vec{w}) = \frac{\sum_{i=1}^{|W|} u_i \cdot w_i}{||\vec{u}_2|| \cdot ||\vec{w}||_2} \quad (3)$$

Coherence provides a well established framework to assess the quality of topic models in an automated way. Besides such an automated analysis, there also exist human-involved methods to validate topics models, e.g., Hoyle et al. [2021], Nanda et al. [2023], Simonetti et al. [2025], and Doogan and Buntine [2021].

Despite its widespread use, coherence has been shown to exhibit limitations. For instance, Hoyle et al. [2021] observed a non-conformity between topics coherence and human evaluations. Nanda et al. [2023] analyze the coverage and the depth of the topics produced via the LDA model. In spite of good results, their findings suggest that the topics require manual interpretation. Doogan and Buntine [2021] claim that the target of interpretability is ambiguous, compromising the validity of both automatic and human evaluation methods.

Simonetti et al. [2025] also state that topic modeling does not guarantee the interpretability of its output and therefore they define a new coherence measure based on co-occurrence networks, where each network represents a topic and its nodes are the most probable words for that topic. In such a network, there is a link between a pair of words if they co-occur in the sentence; thus, the more connected the network, the more coherent the topic in question.

Hoyle et al. [2025] propose to use a prompt-based approach to evaluate topic quality. Their protocol involves comparing the performance of humans and LLM-based proxies in evaluating topic model outputs.

Overall, there is a lack of standardization of evaluation protocols, hampering the analyses and comparisons across different works. The literature includes research that employs a mix of both automated evaluation metrics and human-involved evaluations. In this work we have analysed the Coherence metric, as well as provided a qualitative discussion over the output topics.

2.4 Additional related work and comparisons to our study

Our research is centered on topic modeling and on analysis of inherent latent semantics.

Top2vec [Angelov 2020] infers topic vectors by applying vector algebra over the neural vectors of words and documents embedded in the same vector space. Each topic corresponds to a centroid of a cluster of documents and takes the closest word vectors as the most representative words. The approach infers the optimal number of topics through the HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) algorithm.

Asgari-Chenaghlu et al. [2021] present a model which recognizes topics combining Transformers, community detection techniques in graphs, and named entity recognition.

BERTopic [Grootendorst 2022] is a model based on clusters of documents grouping vectors inferred through the Sentence-BERT framework [Reimers and Gurevych 2019]. It assigns the clusters to a cTFIDF matrix that indicates the most representative words of each topic. cTFIDF is a variation of the classic TFIDF algorithm, taking clusters as the unit of analysis, instead of documents.

Top2Vec and BERTopic rely on HDBSCAN for clustering, which typically requires dimensionality reduction of sentence vectors beforehand. We adopt a Hierarchical Aggregation algorithm instead of HDBSCAN, thereby avoiding dimensionality reduction. Consequently, our method operates on the sentence vectors in their original 768 dimensions, retaining more semantic information. Our approach is similar to that of BERTopic in modeling the topics as clusters of sentences and using cTFIDF to estimate the words' importance in a topic/cluster.

Although many studies use hierarchical clustering algorithms – such as HDBSCAN – their focus is not on the hierarchy, as they apply dimensionality reduction that interferes with hierarchical analysis. We provide an analysis of the inferred hierarchy of topics and its relationship with Language Models based on the Transformer architecture and its attention mechanism.

Our analysis relies on hierarchical topic structures, as explored in Kim et al. [2012], Isonuma et al. [2020], and Yan et al. [2022]. Our approach builds upon Kim et al. [2012], who highlighted important aspects of topic hierarchies and proposed the recursive Chinese Restaurant Process (rCRP) method to generate hierarchical topic structures. We, on the other hand, construct the hierarchy through successive aggregation steps applied to the sentence collection.

Wu et al. [2024] propose a method that uses dependency matrices to model relationships between topics, enabling the construction of topic hierarchies. In contrast, our approach generates topic hierarchies through a simpler process, relying on the successive steps of the clustering algorithm.

Partially similar to our work, Qiu et al. [2024] also present a method that combines SentenceBERT, a clustering algorithm and TF-IDF, but further provide prompt engineering to automatically label the topics.

3 The ABT Method

ABT (Attention-based Topics) is a Neural Topic Model proposed by us to uncover the statistical semantics latent in large corpora. To describe ABT, we formally define the following terms (extending the notation used by Blei et al. [2003] to describe the topic modeling and by Devlin et al. [2019] to define BERT representations):

- Word w: the basic unit of a textual construction. Each w is a discrete data item from the vocabulary $V = \{w_1, w_2, ..., w_v\}$.
- Sentence s: a sequence of n words denoted by $s = (w_1, w_2, \ldots, w_n)$. ABT can work with words, sentences, documents and documents excerpts.
- *Corpus* C: a collection of m sentences denoted by $C = \{s_1, s_2, ..., s_m\}$.
- Topic t: an object characterized by its set of sentences t_S (inferred through step 2) and its set of words t_W (inferred through step 3).
- Representation $\mathbb{R}^{768}$: a multidimensional vector space of 768 dimensions (the standard adopted by BERT-based models) representing the m sentences from *corpus* C. That is, each vector $r \in \mathbb{R}^{768}$ corresponds to a sentence $s \in C$. There are m sentence vectors of 768 dimensions.

3.1 Overview

The ABT pipeline consists of three successive transformation steps. As depicted in Figure 2, our method transforms a given collection of textual sentences C into a set of topics T that encodes the gist of C. It takes C as input in plain-text format and does not require any pre-processing over it.

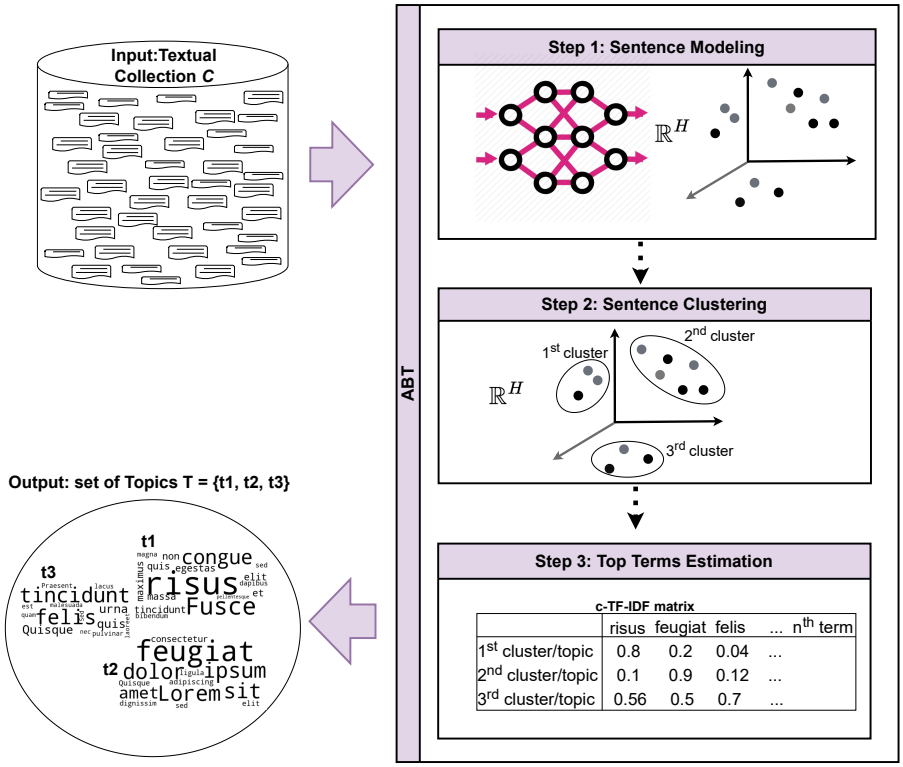

Figure 2: An overview of our 3-step pipeline working with a given collection of sentences. The steps operate over the input collection to produce a set of topics that summarizes it.

Figure 2 illustrates the following 3-steps pipeline:

Step 1: Transform each sentence $s \in C$ into a vector representation $r \in \mathbb{R}^{768}$ using the Sentence-BERT framework [Reimers and Gurevych 2019] to model sentences. The resulting language model incorporates a semantic space of multiple dimensions – usually 768 or 1024, depending on the BERT version.

Step 2: Apply Hierarchical Clustering to $\mathbb{R}^{768}$ (i.e., the output from step 1) to identify a set of disjoint clusters of semantically related sentences. Each cluster characterizes a latent topic t composed by a sentence set t_S.

Step 3: Fill the cTFIDF matrix (proposed by Grootendorst [2022]) with the significance values of each term within the topic. Each row of the matrix corresponds to a topic t characterized by a list t_W containing the most relevant words in the global context of the collection, i.e., recurrent words in the current topic insofar as rare in the other topics.

As output, ABT produces a set of Topics T. Each topic $t \in T$ is characterized by its set of sentences t_S (inferred through step 2) and its set of words t_W (inferred through step 3).

The remainder of this section gives more details of each step of our method using the CliCR corpus [Šuster and Daelemans 2018] as the input collection C of a running example. The CliCR is a corpus of 10,539 Clinical Cases containing descriptions of situations that occur in real clinical environments. Table 1 shows some sentences sampled from the CliCR corpus.

id	Sentence
1	Isolated cranial distortion mimicking caput succedenum from amniotic band disruption without any neurological abnormality
2	Ruptured pseudoaneurysm of the radial artery
3	Pulmonary alveolar microlithiasis
4	Successful pregnancy with autoimmune cirrhosis
...	...
1000	Intraoral schwannoma – a report of two cases
...	...
10539	Thrombotic thrombocytopenic purpura in a patient with HIV from Zimbabwe

Table 1: Textual sentences sampled from the titles extracted from the documents in the CliCR dataset titles of each document.

3.2 Step 1: Sentence Modeling

Step 1 converts the sentences into vector representations embedding them in an H-dimensional vector space $\mathbb{R}^{768}$. This conversion is made by using the Sentence-BERT [Reimers and Gurevych 2019] framework[4] that allows modeling sentences – e.g., paragraphs, documents – using a language model based on siamese BERT-networks. Sentence-BERT infers a neural sentence model which projects the sentences into a vector space of $H = 768$ dimensions (or 1024 depending on the BERT model version) using a base LM.

As a running example of this step, Figure 3 shows the vector space $\mathbb{R}^{768}$ produced by vectorizing the CliCR corpus using the standard BERT model (i.e, bert-base-cased[5]). In fact, this visualization is a 2-dimensional vector space obtained by reducing the dimensionality of $\mathbb{R}^{768}$ from its original 786 dimensions into 2 dimensions (because it is not possible to visualize a vector space of 768 dimensions). We have used a dimensionality reduction technique called Spectral Embedding [Ng et al. 2001] implemented using the Scikit-learn Python module[6].

Sentence-BERT allows one to use any BERT-based language model to produce the vectors, inferring different vector spaces calibrated by specific training tasks. In this section we use the original BERT model as a running example to illustrate the

[4] https://github.com/UKPLab/sentence-transformers
[5] https://huggingface.co/google-bert/bert-base-cased
[6] https://scikit-learn.org/stable/modules/generated/sklearn.manifold.SpectralEmbedding.html

Figure 3: The CliCR corpus projected in a 2-dimensional space (obtained by reducing from the original $\mathbb{R}^{768}$ inferred by the BERT model). Each point represents a clinical case title. In such a space, the closer two points are, the more similar they are

method. As described later in the results section 5, we have experimented with five different BERT-based language models (BERT [Devlin et al. 2019], BioBERT [Lee et al. 2020], Envoy [Pantoja et al. 2022], BART [Lewis et al. 2019], and all-mpnet [Reimers and Gurevych 2019]) to investigate the effects of changing the base model used. Besides the base model, in this step one can configure other hyper-parameters: the model dimensionality (768 or 1024) and the maximal input sentence size (e.g., 128, 512).

3.3 Step 2: Hierarchical Sentence Aggregation

This step involves applying a Hierarchical Aggregation on the set of vectors $\mathbb{R}^{768}$ to obtain clusters of similar sentences. The clustering algorithm known as hierarchical aggregation is an unsupervised method that identifies clusters in a bottom-up approach [Kubat 2017]:

1. initially, each vector $r \in \mathbb{R}^{768}$ defines its cluster;
2. iteratively, the algorithm merges pairs of clusters with the smallest distance from each other among all the cluster pairs;
3. the process continues until it reaches a certain user-defined threshold of distance d – a hyperparameter that indirectly determines the number of output clusters;

4. the algorithm outputs k disjoint clusters corresponding to the set of latent topics T.

In each running of the clustering algorithm, a sentence can only be associated with a single topic. A challenging aspect of topic modeling approaches is the choice of k. An analysis of the clusters hierarchy (as presented in section 5) can help one understand the model's behavior as k increases and the sentences are partitioned into more refined clusters.

Figure 4 illustrates the clustering process over the space produced by the Sentence-BERT model. Figure 4.a presents the dendrogram produced by clustering with $d = 80$ and Figure 4.b illustrates the $k = 3$ resulting clusters.

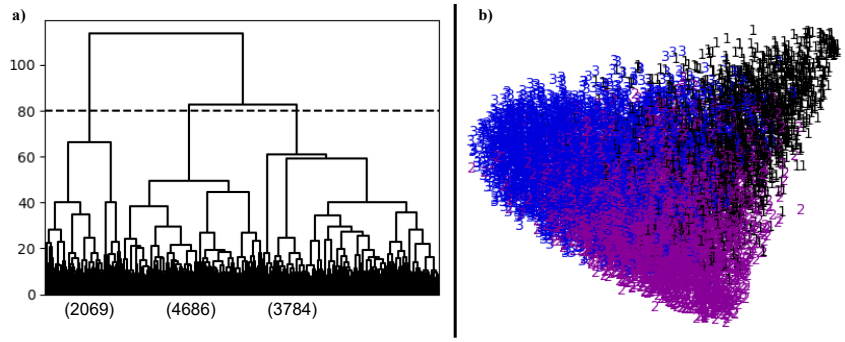

Figure 4: ***a)*** *Dendrogram of the hierarchical clustering of the CliCR corpus modeled by the SentenceBERT model. The hierarchical aggregation algorithm produces 3 resulting clusters by defining the distance threshold $d = 80$.* ***b)*** *Color coded 2d visualization of the 3 clusters of clinical cases. This partitioned space (i.e., the result of step 2) makes explicit the topics encoded in the network parameters.*

Table 2 shows selected samples of clinical case titles and their assigned cluster. The sentences shown in Table 2 were chosen to be representative of the output of this step and to illustrate the similarity within a cluster's sentences. The sentence from a given cluster must be more similar to other sentences within the cluster than those of other clusters. We measured the similarity using the cosine distance between the sentences within the semantic space.

Cluster	Sentence sample
1	Wind direction and mental health: a time-series analysis of weather influences in a patient with anxiety disorder.
	Working effectively with patients with comorbid mental illness and substance abuse: a case study using a structured motivational behavioural approach.
	Ineffective chronic illness behaviour in a patient with long-term non-psychotic psychiatric illness.
2	Isolated splenic tuberculosis in an immunocompetent patient.
	Orbital cellulitis with periorbital abscess secondary to methicillin-resistant Staphylococcus aureus (MRSA) sepsis in an immunocompetent neonate.
	Primary orbital chronic granulomatous reaction to deep staphylococcal infection due to trauma in immunocompetent.
3	Novel balloon application for rescue and realignment of a proximal end migrated pipeline flex embolization device into the aneurysmal sac: complication management.
	Subarachnoid hemorrhage then thrombosis of posterior inferior cerebellar artery dissection: is early surgical exploration warranted?
	Hemorrhagic collision metastasis in a cerebral arteriovenous malformation.

Table 2: Samples of clinical case titles of the 3 clusters produced by the SentenceBERT model. Each cluster gathers a set of similar sentences according to such a specific language model.

The inferred clusters gather sentences that are similar in certain aspects observed, captured, and featured by the language model during its pre-training phase. We notice, for instance, that cluster 3 contains sentences reporting unusual and rare clinical cases. Meanwhile, cluster 1 presents sentences associating symptoms and diagnosis.

It is worth mentioning the special scenario of $k = 1$ (i.e., a single topic to represent the entire collection). In such a case, it is not possible to assemble the cTFIDF matrix since it requires at least 2 clusters. Therefore, to create a $k = 1$ topic the ABT uses a classical TFIDF matrix, considering each input sentence $s \in C$ as a row, which implies a slightly different topic than those $k > 1$ created with the cTFIDF matrix.

The resulting clusters form the basis of the semantic representation built by ABT. We are interested in investigating whether there is a task well suited for topic modeling and apt to produce an optimal set of clusters. Therefore in the Results Section 5, we experiment with different threshold distances to infer a range of partitions of the input data set.

3.4 Step 3: Representing Topics

Lastly, ABT obtains the most descriptive terms from each cluster – through the cTFIDF measure [Grootendorst 2022], which indicates the relevance of each term within the cluster – and assigns them to the corresponding topic $t \in T$, according to:

$$cTFIDF_{w,t} = TF_{w,t} \times \log(\frac{k}{TF_w}) \tag{4}$$

Where $TF_{w,t}$ models the frequency of the word w in the topic t. Here, the topic t is the concatenation of all sentences s in the cluster that corresponds to t. The Inverse Topic Frequency IDF measures how much information a word w provides to a topic t by taking the logarithm of the total number of topics k divided by TF_w the frequency of word across all topics. In other words, TF_w is the number of topics that contain w.

Then, it fills the cTFIDF matrix by allocating the cluster in the rows and the vocabulary words $V = \{w_1, w_2, ..., w_v\}$ in the columns, as exemplified in Table 3. The cTFIDF algorithm gives information on how relevant each term is to the referred cluster, considering the common terms to the cluster insofar as uncommon in other clusters.

Cluster	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10	...	Word 7320
1	mental 0.14	guillain 0.12	your 0.12	still 0.10	illness 0.10	kikuchi 0.09	barr 0.08	this 0.08	about 0.08	forget 0.08	...	$\beta3$ 0.0
2	immunocompetent 0.26	cardiomyopathy 0.19	cystic 0.15	diffuse 0.15	mutation 0.13	intravenous 0.12	month 0.11	autoimmune 0.10	chemotherapy 0.10	staphylococcus 0.10	...	laptop 0.0
3	aneurysm 0.47	dissection 0.28	arteriovenous 0.19	gastrointestinal 0.18	ruptured 0.14	occlusion 0.14	inferior 0.12	uterine 0.12	vertebral 0.12	maxillary 0.11	...	$\beta3$ 0.0

Table 3: Example of cTFIDF matrix to 3 topics inferred by using the SentenceBERT model. Each topic contains a list of top words sorted by cTFIDF metric. For example, the top word "mental'' from Cluster 1 exhibits cTFIDF = 0.14.

The output produces a set T containing a total of k topics, where each topic $t \in T$ is characterized both by the set t_W (containing its most discriminant words) and by the set t_S (containing its clustered sentences).

Figure 5 shows the topics produced in the output of this step, represented in the vector space by color, and, for each, we present a word cloud with the best-ranked terms in cTFIDF.

This step allows one to filter out stop-words from the corpus by setting the MAX_TF IDF hyper-parameter, which filter outs words that are frenquently above this threshold.

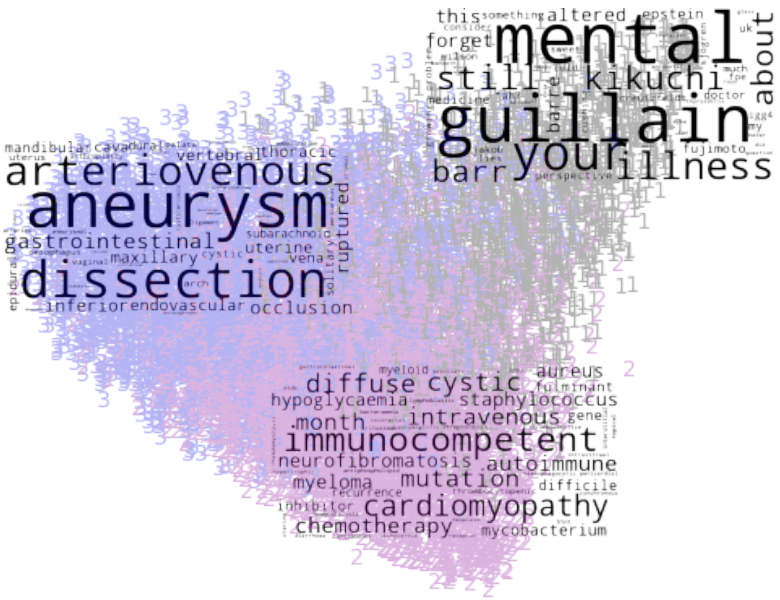

Figure 5: Word clouds of most relevant terms by each topic, inferred by using the ABT_{BERT} model.

In this way, the method dynamically infers a list of stop words, thus alleviating one of the bottlenecks inherent in classical approaches that use fixed and predefined stop-word lists [Nothman et al. 2018].

4 Materials and Methods

We have applied a type of probing methodology [Tenney et al. 2019] to investigate whether ABT consistently models some (if any) information associated with the topic model's phenomena. We have conducted five distinct analyses:

1. a quantitative analysis regarding the Topics Coherence metric.
2. an analysis of the the generalizability of the proposed ABT model in different domains, by applying ABT to two distinct data collections: CliCR and MovieLens.
3. a qualitative analysis of the Hierarchy Structure underlying the set of discovered topics.
4. an ablation study arguing about the misalignment between the metric Topics Coherence and human evaluations.

5. an analysis of the computational cost of ABT compared to LDA.

We have analyzed the use of five BERT-based language models (BERT [Devlin et al. 2019], BioBERT [Lee et al. 2020], Envoy [Pantoja et al. 2022], BART [Lewis et al. 2019], and all-mpnet [Reimers and Gurevych 2019]) in the first step of the pipeline, run on a Linux PC with 16 GB of RAM size, Intel i7, programmed in Python.

The experiment setting explored an exhaustive search of the following hyperparameters:

- MAX_TFIDF thresholds: $\{0.001, 0.01, 0.1, 0.9, 0.99, 1.0\}$;
- Number of topics $k = \{1, 2, 3, ..., 200\}$;
- Number of top words per topic $n_w = 10, 20, 50$.

We have analyzed the performance of our model in two real-world datasets: CliCR [Šuster and Daelemans 2018] and MovieLens [Harper and Konstan 2015]. CliCR is a collection of Clinical Cases, where each case describes a medical narrative of situations occurring in real clinical environments. MovieLens consists of movie records containing titles along with their genres.

We have chosen CliCR because clinical cases are important educational/medical resources. In medical education, professors use clinical cases as pedagogical resources to teach clinical practices to medical students. We have chosen MovieLens because it is widely used in education and research. Also, movies are of common interest, making the results easy to interpret and present.

As a proof of concept, our analysis focused on datasets of short pieces of texts (see Table 4) to show the effectiveness when dealing with small lengths of text. However, ABT is flexible enough to handle larger text fragments – by improving the Sentence-BERT model to deal with more characters than the standard 128 BERT-tokens. Although such an improvement is theoretically feasible, it would demand more computational resources (e.g., more memory and maybe the usage of GPUs).

	CliCR	**MovieLens**
Number of documents	10,539	9,742
Avg. documents title length (in characters)	79.12	25.85
Avg. documents length (in characters)	1457	–

Table 4: The CliCR and MovieLens corpus statistics.

5 Evaluation of Results

5.1 CliCR

The analysis of the CliCR dataset focused on case titles rather than the full content of clinical cases (see Table 4) as a proof of concept. Although theoretically feasible, processing complete case narratives would demand greater computational resources.

Table 5 shows the quality in terms of coherence of some (among fifty) topics produced in an experiment round of ABT with $k = 50$. For instance, the best-ranked topic (row 1)

exhibits coherence score $c = 0.59$ with some top words as {old, year, in, ...}, while the worst topic (last row) achieves $c = 0.11$ with some top words as {syndrome, in, davidoff, ...}. These discriminant words may indicate the emergence of a specialization pattern as coherence decreases, under the specific hyperparameter setting MAX_TFIDF= 0.001.

		Top Words									
Topic	**Coherence**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	0.59	old	year	in	boy	with	man	and	an	girl	woman
2	0.57	the	in	an	and	with	case	syndrome	to	or	is
3	0.56	rare	tumour	carcinoma	the	presentation	an	cell	cyst	with	unusual
4	0.55	obstruction	in	the	dissection	bowel	an	gastrointestinal	artery	to	malformation
5	0.55	in	with	patient	and	an	the	syndrome	following	to	case
...											
11	0.51	in	staphylococcus	with	streptococcus	infection	aureus	and	mycobacterium	methicillin	difficile
12	0.50	intracranial	adrenal	neurofibromatosis	with	in	autoimmune	neuropathy	hypertension	type	and
13	0.50	the	in	an	and	with	schwannoma	hernia	en	case	to
14	0.50	cause	an	unusual	rare	pain	in	disease	case	acute	common
15	0.50	the	cyst	in	an	with	gallstone	compound	hydatid	and	odontoma
...											
36	0.43	the	artery	left	coronary	cruciate	ligament	rupture	circumflex	tendon	anterior
37	0.42	in	intussusception	urinary	the	vaginal	neonate	rupture	pregnancy	preterm	an
38	0.40	with	in	and	the	patient	pancreatitis	an	thrombocytopenic	acute	as
39	0.40	mandibular	the	maxillary	supernumerary	mandible	molar	palate	in	lateral	incisor
40	0.40	the	an	mass	body	unusual	foreign	masquerading	as	in	nose
...											
46	0.35	fracture	dislocation	the	fixation	in	injury	and	with	an	anterior
47	0.35	volvulus	meckel	calcification	diverticulum	an	schonlein	the	in	henoch	lichen
48	0.35	in	and	osteogenesis	with	imperfecta	bilateral	herpetic	an	fistula	spontaneous
49	0.33	syndrome	sarcina	tuberculosis	disease	mclean	isaacs	brokenosis	glassidium	di	sala
50	0.11	syndrome	in	davidoff	dyke	masson	guillain	hauser	küster	creveld	ellis

Table 5: Some example topics in descending sort of coherence scores. The most relevant terms describe each topic according to the cTFIDF metric.

Table 6 relates the number of topics with the coherence obtained by using a selected set of language models: BERT, BioBERT, Envoy, BART, and all-mpnet. We have tested various neural language models to investigate whether using them in ABT's first step affects the result, as each language model outputs a set of topics which gathers similar sentences according to the corresponding model. We are interested in observing the behavior of our method by applying language models calibrated by different training tasks: whether better-suited tasks produce more meaningful sets of topics.

Each row in Table 6 shows the coherence achieved by ABT when experimenting with the models standing in the columns. For instance, in the first row: all the models exhibit a similar coherence score $c = 0.60$ when producing 1 topic. By requesting $k = 50$ topics, the models show different performances, (e.g., BioBERT achieves $c = 0.45$ and Envoy $c = 0.46$). BioBERT scores $c = 0.38$ when producing $k = 200$ topics.

These quantitative experiments showed that although models' performance is consistent across different language models, none completely outperforms the others regarding coherence results. However, if we look at the results, we can see that some topics are more (or less) informative than others, as depicted by Figure 6, which shows the topics produced by ABT (by working with BERT, Envoy, BioBERT, all-mpnet and BART models) and by LDA. The columns present the topics learned by using the models listed in the corresponding rows. We set the hyperparameter MAX_TFIDF to a fixed value of 0.001 for both the LDA and ABT variants. The output topics are substantially different from one another.

In the comments that follow, we consider a word as *plain language* if its meaning can be understood by anyone. Conversely, *specific words* are those that are understood only by domain experts.

It is possible to observe topics produced by LDA that have plain language words scattered throughout (e.g., {unusual, case, patient, rare, woman, following}). On the other hand, ABT produces topics that are more domain-specific than LDA, as well as

	Topic Coherence by Model					
k	**ABT$_{BERT}$**	**ABT$_{Envoy}$**	**ABT$_{BioBERT}$**	**ABT$_{BART}$**	**ABT$_{all-mpnet}$**	**LDA**
1	0.60	0.60	0.60	0.60	0.60	0.60
2	0.56	0.57	0.54	0.54	0.55	0.58
3	0.56	0.59	0.58	0.52	0.55	0.57
4	0.53	0.56	0.55	0.52	0.56	0.58
5	0.55	0.57	0.53	0.52	0.56	0.56
6	0.55	0.54	0.53	0.54	0.56	0.56
7	0.54	0.55	0.52	0.53	0.55	0.55
...						
50	0.45	0.46	0.45	0.45	0.45	0.36
...						
200	0.39	0.39	0.38	0.39	0.37	0.31
...						
10539	0.42	0.42	0.42	0.42	0.42	0.31

Table 6: Topics Coherence results achieved both by ABT variants and classical LDA. According to the scores, ABT is on par with LDA.

Model	**Topic 1**	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**	**Topic 6**
ABT$_{BERT}$	bilateral aneurysm spontaneous	always it use	report carcinoma associated literature	pregnancy treatment two	boy year	treatment
ABT$_{Envoy}$	foreign use	drugcardiomyopathy encephalopathy infarction hepatitis	squamous metastases metastasis sarcoma	intussusception malformation diverticulum myeloid	endocarditis infarction arthritis septic pneumonia	neurofibromatosis metastasis
ABT$_{all-mpnet}$	coronary cerebral aortic carotid infarction	poisoning foreign	immunocompetent endocarditis tuberculosis lupus encephalitis	thyroid adrenal sarcoma	hernia fracture dislocation	fracture
ABT$_{BiobERT}$	it not two or	secondary report pregnancy	secondary leukaemia haemorrhage lymphoma	squamous adenocarcinoma metastasis	aneurysm dissection myocardial infarction	boy month
ABT$_{BART}$	patient rare	presenting patient	beyond tale nostras	patient rare	patient artery	year
LDA	medullary dasatinib reversed	control proven tremor hyperthyroidism	burns mucosa spondylodiscitis branch	excision testis malaria	globe closed stridor neglected	calcium calcific

Figure 6: Topics representing CliCR dataset.

some plain language ones. For example, the ABT based on BART (ABT$_{BART}$ at the 5^{th} row of the figure) produces topics containing plain language words (e.g., {patient, presentation, report, month}).

The ABT$_{envoy}$ and ABT$_{all-mpnet}$ models contain fewer plain language words than the remaining models (e.g., {use, approach, foreign, care}). Thus, ABT can be configured to use fine-tuned LMs to produce topics that are more domain-specific than LDA.

In these topics, it is possible to visualize some evident patterns, such as recurring concepts – e.g., "endocarditis" in ABT$_{Envoy}$ and $ABT_{all-mpnet}$, "aneurysm" in ABT$_{BERT}$ and ABT$_{BioBERT}$, and "metastasis" in ABT$_{BioBERT}$ and ABT$_{Envoy}$.

It is worth commenting on the topics produced by the ABT$_{Envoy}$ model. For instance, almost all the topics seem to allude to anatomies, chemicals or diseases, except for topic

{foreign, technique, approach...} which stands for more generic terms than the other topics (although also related to the biomedical context). We suppose that these topics arise as a natural reflection of the pre-training objective of the Envoy model, which consists in recognizing 4 different types of named entities: Chemical, Anatomy, Disease and Other. Therefore, the Envoy model deems the presence of terms alluding to such entities as discriminant features to infer the space vector $\mathbb{R}^{768}$ to represent the CliCR corpus. In this sense, such more generic topic could be elicited from a region in the vector space crowded with sentences containing generic terms classified as Other. Conversely, all the other topics gather terms alluding to anatomies, chemicals or diseases.

In order to observe the evolving process which produces topics, in Figure 7 we have organized a partial depiction of ABT$_{BERT}$'s topics in a tree structure of 7 levels. We refer to each topic by its level on the hierarchy added by its index in the level. For example, the root topic $t_{1,1}$ is the topic produced by ABT when asked to produce $k = 1$, i.e., one single topic. Similarly, $t_{3,1}$ ={mental, guillain, your, ...} refers to topic 1 at level 3. Each level displays the topics obtained on the k^{th} round of ABT.

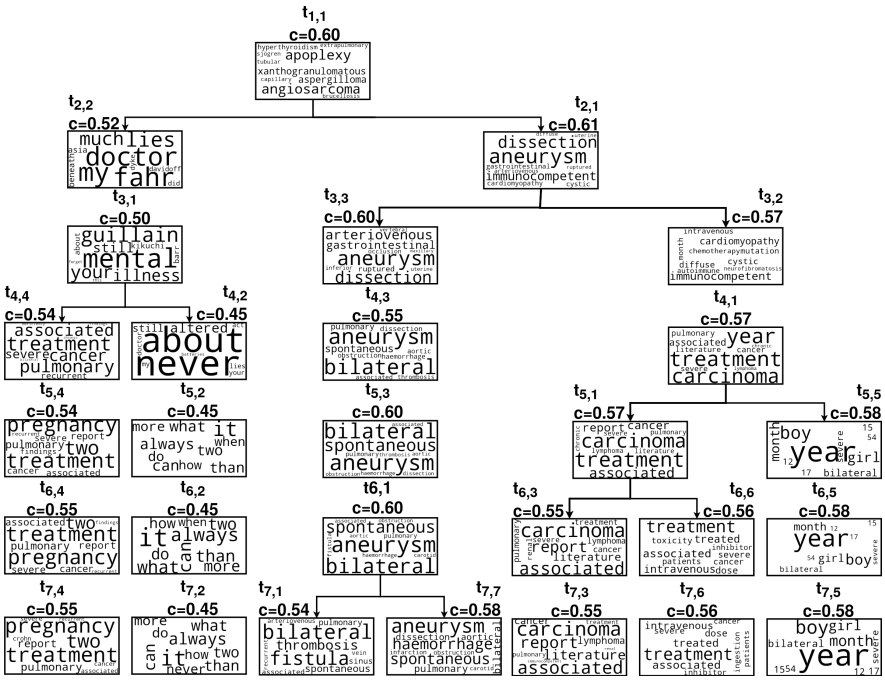

Figure 7: The top of the hierarchical structure produced through ABT$_{BERT}$. It is possible to observe some interesting patterns – such as, topics in the leaves are more specific than those near the root and the coherence gets low as the hierarchy goes down.

The ABT topic hierarchy exhibit the Hierarchical Affinity characteristic [Kim et al. 2012]: topics that descend from a topic t must be more similar to t than topics descending from other topics. For example, topic $t_{2,1}$ ={aneurysm, dissection, immunocompetent, gastrointestinal, cystic, cardiomyopathy, arteriovenous, ...} is more similar to its children

topics $t_{3,3}$ ={aneurysm, dissection, arteriovenous, gastrointestinal, ruptured, ...} and $t_{3,2}$ ={immunocompetent, cardiomyopathy, cystic, diffuse, mutation, intravenous, ...} than to its non-children topic $t_{4,2}$ ={never, about, altered, still, your, my, ...}.

Furthermore the hierarchy is consistent with the property of non-monotonicity [Röder et al. 2015] in which the coherence of a given topic is increased by adding related sentences and must decrease by adding non-related sentences. In our approach we have analysed how the coherence behaves in a growing set of topics. In overall the coherence appears to decrease as the hierarchy goes down (that is, k increases). However, some split points seems to break such monotonic behaviour. For example, split of topic $t_{3,1}$ give rises to the topics $t_{4,4}$ and $t_{4,2}$, the first one has an increase in coherence from $c = 0.50$ to $c = 0.54$ and, oppositely, $t_{4,2}$ decreases from $c = 0.50$ to $c = 0.45$. It can indicate that the parent topic $t_{3,1}$ has gathered sentences not so many related and therefore break the monotonic growth.

The monotonic behaviour is apparent in the branch constituted by the topics $t_{1,1}$, $t_{2,2}$, $t_{3,1}$, $t_{4,2}$, and $t_{5,2}$. This sequence of topics are changing various times as the hierarchy goes down, until to reach a stable point at $t_{5,2}$ ={it, always, ...} mainly constituted by plain-language words. This branch instability maybe indicate that theirs topics does not represent the collection at all and could indicate outliers in the input data.

Our approach can indicate the balance of a topic when it is split into subtopics. Let us consider two terms of similar prevalence in topic $t_{6,1}$: “immunocompetent” and “aneurysm”. The term “immunocompetent” seems to have a more balanced distribution between topics $t_{7,7}$ and $t_{7,1}$, dividing the force of its prevalence, and incurring in a increase of coherence when joining those topics.

Despite the satisfactory performance achieved on coherence scores, we observed that the hierarchy structures produced did not hold the generic-to-specific property (Topic specialization) elicited by Kim et al. [2012] and Isonuma et al. [2020]. According to the property of topic specialization, the topics close to the root must be more generic than the leaf topics. Our topic hierarchy does not always follow this property – sometimes, generic topics appear closer to the leaves than to the root. We suppose that, to generate strictly hierarchical structures, our method would demand a sentence model (in step 2) that discriminates sentences according to their generality and specificity. For example, one could experience with a hypernym-hyponym detection task to better separate the sentences in the vector space according to their generality degree. Such a task could be used to infer general-to-specific topic hierarchies. However, we suspect that a perfect separation would not be possible since generality and specificity are subjective matters, which would imply the need for human intervention - and potentially more than one hierarchy alternative.

5.2 MovieLens

MovieLens [Harper and Konstan 2015] is a dataset extracted from the MovieLens online recommendation system. The dataset consists of 9,742 movie records containing (in addition to other data) movie titles along with zero or more genres, for example (A Bug’s Life | Animation | Children’s).

We have used MovieLens to get a direct comparison with the well established work of Kim et al. [2012], in order to fill a gap of comparative analysis on the evaluation protocols commonly employed by the literature.

Kim et al. [2012] applied the rCRP (Recursive Chinese Restaurant Process) topic modeling approach for discovering hierarchical topic structures. Part A of Figure 8 is a picture reproduced from their paper, while Part B shows our results from produced by our

$ABT_{all-mpnet}$ method. The picture indirectly reflects semantics of topics by exhibiting the movies which better represent the genre of each topic. For example, the topic Horror is characterized by the movies "The Exorcist, The Shining, Halloween, ...". In their work, topics are labeled by experts, whereas in Part B we suggest using word clouds to provide an initial label for a topic, since ABT does not consider expert intervention.

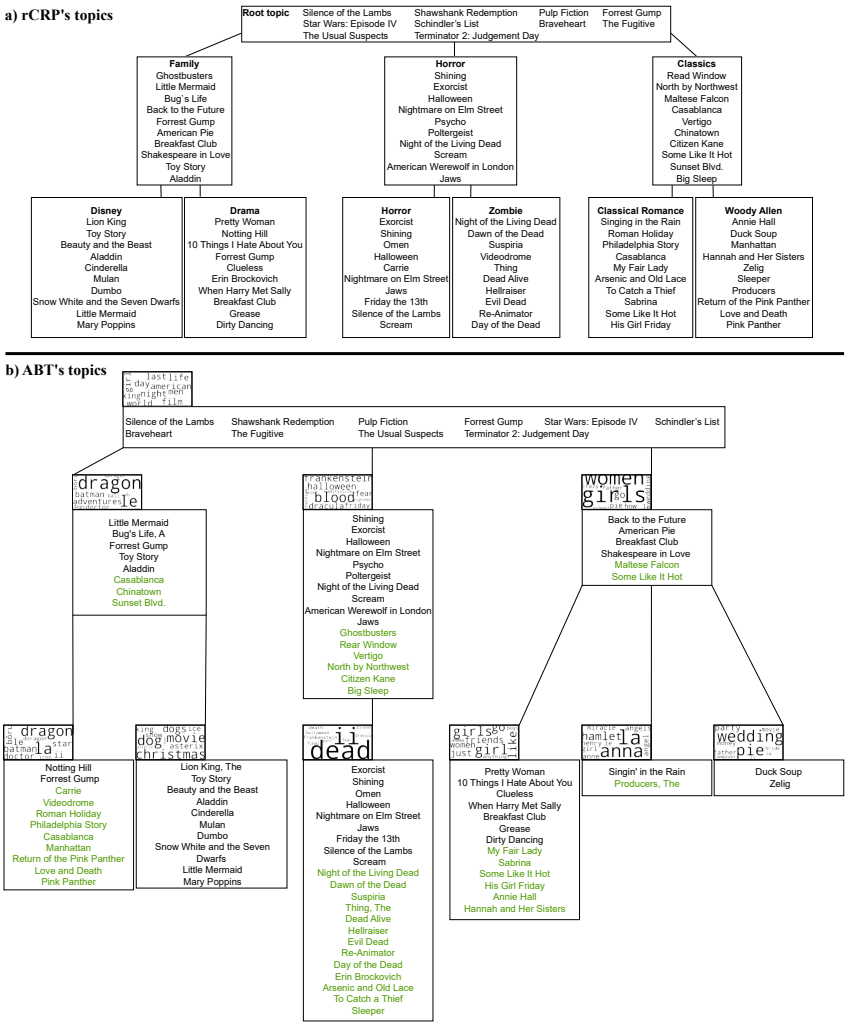

Figure 8: A comparative chart between ABT's topics and topics presented by Kim et al. [2012]. The topics are represented by sampled sentences.

Besides these labeling differences, it is possible to observe in the Figure similarities and differences between rCRP and ABT output topics. First of all, the topics at the

root of each tree are the same, but they begin to diverge afterwards. Movies colored with green represent movie titles that were classified in different topics by our model. For example the movie CasaBlanca was classified by rCRP within the topic Classical Romance, whereas in our model it is classified in the topic containing the movies {Litte Mermaid, A Bug's life, ...}, a misalignment between rCRP and ABT models.

There is, on the other hand, a full alignment between the topic Disney (from rCRP) and the topic containing the movies {The Lion King, Toy Story, ...}. On the other hand, the topics Horror and Zombie (from rCRP) correspond to one single topic in ABT topics {The Exorcist, The Shining, ...}.

While rCRP requires human annotators to label topics, ABT elicits the word clouds for the same goal. The wordclouds in the figure help one to gain insights about the topic semantics by comparing the sampled movies with the main words of that topic. Future work can investigate metrics to measure how well a wordcloud represents a topic, and its constituent documents.

5.3 Computational Cost

Table 7 shows the memory usage and runtime execution comparing ABT against LDA. The costs of ABT are split into steps to highlight the contribution of each step.

	k	ABT			LDA
		Step 1	Step 2	Step 3	
Memory Usage (GB)	1	1.45	0.01	0.15	0.15
	10	1.45	0.19	0.1	0.22
	100	1.63	0.01	1,01	0.23
	1000	1.88	0.01	0.86	0.31
	10539	1.86	0.01	0.87	1.08
Running Time (seconds)	1	238.75	14.32	0.57	1.68
	10	246.53	12.74	0.24	1.52
	100	222.86	13.06	8.62	2.05
	1000	288	13	9	6.21
	10539	243.57	13.30	8.76	59.42

Table 7: Cost comparison between ABT and LDA.

According to Table 7, ABT is memory-intensive and takes a long time to run. Despite these drawbacks, its topics are more coherent than those produced by LDA (as previously shown in Table 6). Thus, there is a trade-off between computational cost and performance: LDA is more computationally efficient, whereas ABT achieves higher coherence scores.

ABT steps 2 and 3 are lightweight compared to step 1, which is memory-intensive and time-consuming. This makes it possible to configure ABT for real-time applications: step 1 is loaded once at application startup, while only steps 2 and 3 may be repeatedly executed in response to user requests for new topics. Step 1 would be invoked again only when a new sentence must be added to the vector space or when one wants to change the base LM.

For small values of k (the number of topics), the cost of LDA is much lower than that of ABT. However, as k increases, LDA requires more memory and longer running

times, whereas ABT maintains an almost constant use of these resources. Thus, in terms of scalability with respect to the number of topics, ABT is well suited for large-scale applications.

5.4 Ablation Study

Our results show that ABT's topics are at par with the ones produced by the traditional method LDA, and the most recent neural topic models.

Despite these results, the topic coherence that we measured did not correlate with the semantics of a topic. For example, topics t_{11} ={in, staphylococcus, with} and t_{50} ={syndrome, in, davidoff} from Table 5 have very different coherence values (0.51 and 0.11, respectively) and yet it is not easily seen which one "is the best". According to our investigation studies, this misalignment could be associated with an issue already observed and addressed by Hoyle et al. [2021]. This will be examined next.

Hoyle et al. [2021] call "into question the validity of fully automatic evaluations independent of human judgments". They compared the human judgments against automated metrics and found that there is a discrepancy between these two evaluation approaches, that is, "the inherent uncertainty of human judgments means that it is difficult to determine when an increase in a model's mean automated coherence implies a significant improvement in the corresponding human scores".

We argue that the non-conformity between topics coherence metrics and human evaluations – as originally claimed by Hoyle et al. [2021] and reinforced by our results – is a reflection of the impossibility for a language model to learn semantic meaning by training it just on linguistic form (defined as, "any observable realization of language: marks on a page, pixels in a digital representations of text, or movements of the articulators" [Hoyle et al. 2021]). In theory, such impossibility would be grounded on the linguistic concept of meaning as the relationship between a linguistic form and communicative intent, which in overall is something external to language [Bender and Koller 2020] (i.e., not present within the training dataset). Therefore, communicative intents would be outside the scope of the language itself, needing to be added by those tuning a LLM. However, even inputting unlimited numbers of forms (and thus unlimited extra information), a language model would still be unable to learn the complete meaning which is conceived by the language [Bender and Koller 2020]. This constitutes a fundamental limitation for LLMs: even in a scenario where unlimited data are continuously provided, a language model could never fully grasp all the meanings conveyed by text, because essential aspects of meaning lie outside the training data.

Starting from this concept of meaning assumed by Linguistics, Bender and Koller [2020] propose the Octopus Test to illustrate the challenges of learning meaning from form alone. According to their conceptual experiment, a language model trained purely in written texts cannot pass the Octopus Test because it is not capable, a priori, of capturing all the relevant features needed to produce utterances aligned with the evaluations of human judgments, full of semantic meaning. Hence, neural language models inevitably would fail the Octopus Test, and consequently in the Turing Test, since the Octopus Test is a weak form of the Turing Test.

Therefore, a topic model based on neural language models does not capture fundamental features to infer semantics because during training it is exposed only to linguistic form. Following these premises, we assume that the topics coherence matrics may not fully correlate with the human-made evaluations, because the semantics of a topic are subjective matters that depend on "fitness-for-use", demanding human assumptions

constructed from the human mental "database", which is totally different from the binary database available to the text-based coherence evaluator. The mental database of the human evaluator has access to information (such as factual knowledge, intention, memories of smell, touch, social interactions, etc) beyond that contained in the data set used to train the language model.

The ability to acquire context-based semantics requires human interaction and validation. Scholarly work on language acquisition suggests that "the process of acquiring a linguistic system relies on joint attention and intersubjectivity: the ability to be aware of what another human is attending to and guess what they are intending to communicate" [Bender and Koller 2020]. In other words, human children do not learn meaning from form alone. Hence we should not expect machines to do so either [Bender and Koller 2020], as LMs may pick up idiosyncratic and superficial patterns in the data, which are not necessarily "meaning". As a consequence, a neural topic model trained just on textual form will not learn meaningful topics perfectly aligned with human evaluations that also embed semantics. Indeed, topic modeling demands the ability to ground language in the world, which has yet to be fully met by AI research, given that semantic similarity between text fragments is "a potentially useful, but incomplete, reflection of the actual meaning" [Bender and Koller 2020].

This limit imposed on machine learning methods to produce semantic topics does not negate the utility of topic models and language models. Indeed, topic modeling may improve many important NLP tasks. A particular case is the use of topic modeling as a dimensionality reduction technique. ABT can be seen as a summarization framework that reduces the asymptotic complexity of a collection to k dimensions – i.e., k topics – coherent with the successive partitions performed throughout the clustering process. Future analyses can investigate how much discriminatory information it loses by reducing the collection to these dimensions.

6 Conclusions and Future Work

The topic modeling taskms to identify predominant themes in extensive textual collections. It has the potential to improve services and systems that work with textual data. This paper presents our approach to this problem.

ABT treats a topic as a dense region of sentences coupled with semantics, characterized by its most discriminant terms, benefiting from the statistical context emerging from the structure (the vector space) inferred by the neural network. The vector space represents semantics associated with the gist [Griffiths et al. 2007] of the sentence collection, extracted through a statistical inference performed by a neural language model. Grounded on this concept of semantics, ABT organizes topics in a hierarchical tree, which alludes to the semantics of a massive collection of sentences; such semantics are shared by the collective consciousness of co-authors of these texts.

We have reported on experiments using a range of neural language models, run on the CliCR corpus. Our study shows that the proposed approach adequately fits the problem and produces results comparable to the state-of-the-art concerning topics coherence metrics. Our discussion of results analyzes claims that neural language models cannot, in principle, learn the linguistic meaning, a vital component to complete the topic modeling task successfully and also to produce generic-to-specific hierarchies. Topics coherence metrics, consequently, do not align with the topics produced and/or evaluated by humans. Additionally, we analyzed the hierarchical topic structure built as a by-product of the inferring topics process. The coherence achieved at each level of such a hierarchy structure

may indicate the most appropriate levels of the hierarchy to be chosen to summarize a textual collection.

We believe that future explorations, with more grouping levels, may elicit more topics that can be intuitively associated with themes and possibly choose those levels that group topics in the desired way.

In addition, the semantic representations incorporated into attention-based topics can leverage information search systems. Moreover, topic models can be used to assess an LM's performance with respect to a given set of problems.

Additionally, we intend to formalize ABT and analyze its complexity. We plan to integrate our method into pipelines of other more complex tasks. Moreover, we intend to extend our tests to conduct analyses of the method's behavior when working with the full content of clinical cases and with textual collections from other domains. Furthermore, validation can use other metrics – such as Perplexity. Another extension would be to experiment with datasets containing human-annotated rationales to enhance the quality of topics.

Moreover, we intend to include a small-scale human evaluation of the topics generated, to further support our claim of the misalignment of this kind of evaluation with automated coherence metrics. We argue that this may be a consequence of the semantics of a topic being of subjective matters, demanding human assumptions constructed from the human mental "database", which is totally different from the binary database available to the text-based coherence evaluators. As a consequence, neural topic models, trained just on textual form, will not learn meaningful topics perfectly aligned with human evaluations.

Acknowledgements

This study was financed in part by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil* (CAPES) - Finance Code 001 and by the *Fundação de Amparo à Pesquisa do Estado de São Paulo* (FAPESP) - Project 2013/08293-7, and by the *Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brasil* (CNPq) projects #308018/2021-4 and #428459/2018-8.

We thank the *Universidade Federal do Pará* for also supporting this research.

A Appendix

This section presents the Python codes for the pipeline of ABT (A.1) and for the validation by topic coherence (A.2). ABT is an alternative software-based approach to the topic modeling problem. The ABT framework consists of software modules (i.e., steps 1, 2 and 3) that work in tandem and can be freely parameterized and tuned.

A.1 Pipeline

```
DATASET_PATH = 'clicr/train1.0.json'

corpus = []
topics = []

class Topic:
```

```
        def __init__(self, index):
                self.words = []
                self.sentences = ""
                self.coherence = 0.0
                self.index = index

##########
### DATA LOADING AND TITLE EXTRACTION
import json

with open(DATASET_PATH) as f:
        dataset = json.load(f)

        for datum in dataset["data"]:
                title = datum["document"]["title"].replace("BEG__", "").
                    replace("__END", "")
                if (title != 'This article has a correction'):
                        corpus.append(title)

##########
### STEP 1
from sentence_transformers import SentenceTransformer

BERT_MODEL = "bert-base-cased"
language_model = SentenceTransformer(BERT_MODEL)
vector_space = language_model.encode(corpus)

##########
### STEP 2
DISTANCE_THRESHOLD =300

from sklearn.cluster import AgglomerativeClustering

clustering_model = AgglomerativeClustering(distance_threshold=
    DISTANCE_THRESHOLD, n_clusters=None, linkage="ward")
clustering_model = clustering_model.fit(vector_space)
k = clustering_model.n_clusters_

topics = [Topic(i) for i in range(k)]

for i in range(len(corpus)):
        inferred_cluster_index = clustering_model.labels_[i]
        topics[inferred_cluster_index].sentences+= corpus[i]+" "

##########
### STEP 3
MAX_TFIDF = 0.001

import numpy as np
from sklearn.feature_extraction.text import TfidfVectorizer

if (k > 1):
        c_MAX_TFIDF = 0.5
        c_tfidf_model = TfidfVectorizer(max_df=c_MAX_TFIDF)
        c_tfidf = c_tfidf_model.fit_transform([topic.sentences for topic
            in topics])
        c_tfidf_matrix = c_tfidf.toarray()
        words = c_tfidf_model.get_feature_names_out()

        for i, topic in enumerate(topics):
```

```
            sorted_term_indexes = np.argsort(-1*c_tfidf_matrix[topic.
                index])
            topic.words = [words[j] for j in sorted_term_indexes]
else:
       c_tfidf_model = TfidfVectorizer(max_df=MAX_TFIDF)
       c_tfidf = c_tfidf_model.fit_transform(corpus)
       c_tfidf_matrix = c_tfidf.toarray()
       words = c_tfidf_model.get_feature_names_out()

       mean_tfidf = np.array(c_tfidf_matrix.mean(axis=0)).flatten()
       sorted_term_indexes = np.argsort(-1*mean_tfidf)
       topics[0].words= [words[j] for j in sorted_term_indexes]
```

A.2 Validation

```
TOP_WORDS = 10

from gensim.models import CoherenceModel
from gensim.corpora import Dictionary

tfidf_model = TfidfVectorizer(max_df=MAX_TFIDF)
tfidf = tfidf_model.fit_transform(corpus)

words_by_sentence = tfidf_model.inverse_transform(tfidf)

dictionary = Dictionary(words_by_sentence)
cm = CoherenceModel(topics=[topic.words for topic in topics], texts=
    words_by_sentence, dictionary=dictionary, coherence="c_v",topn=
    TOP_WORDS)

coherence_per_topic = cm.get_coherence_per_topic()

for i, coherence in enumerate(coherence_per_topic):
       topics[i].coherence = coherence
       print("Topic ", str(i))
       print(topics[i].words[:TOP_WORDS])

print("Total coherence: ", topics_coherence)
print("Coherence by topic: ", cm.get_coherence_per_topic())
```

References

[Abdelrazek et al. 2023] Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., Hassan, A.: "Topic modeling algorithms and applications: A survey"; Information Systems, 1-17.

[Angelov 2020] Angelov, D.: "Top2vec: Distributed representations of topics"; arXiv preprint arXiv:2008.09470v1, 1-25.

[Asgari-Chenaghlu et al. 2021] Asgari-Chenaghlu, M., Feizi-Derakhshi, M., Farzinvash, L., Balafar, M., Motamed, C.: "TopicBERT: A cognitive approach for topic detection from multimodal post stream using BERT and memory–graph"; Chaos, Solitons & Fractals, 1-13.

[Bender and Koller 2020] Bender, E., Koller, A.: "Climbing towards NLU: On meaning, form, and understanding in the age of data"; Proc. 58^{th} annual meeting of the Association for Computational Linguistics, 5185-5198.

[Bengio et al. 2000] Bengio, Y., Ducharme, R., Vincent, P.: "A Neural Probabilistic Language Model"; Advances in Neural Information Processing Systems, 1137-1155.

[Blei et al. 2023] Blei, D., Ng, A., Jordan, M.: "Latent dirichlet allocation"; Journal of Machine Learning Research, 993-1022.

[Bojanowski et al. 2017] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: "Enriching word vectors with subword information"; Transactions of the Association for Computational Linguistics, 135-146.

[Bouma 2009] Bouma, G.: "Normalized (pointwise) mutual information in collocation extraction"; Proc. German Society for Computational Linguistics, 1-11.

[Boyd-Graber et al. 2014] Boyd-Graber, J., Mimno, D., Newman, D.: "Care and feeding of topic models: Problems, Diagnostics, and Improvements"; 1^{st} ed.; Chapman and Hall/CRC, 225-254.

[Boyd-Graber et al. 2017] Boyd-Graber, J., Hu, Y., Mimno, D.: "Applications of topic models"; Foundations and Trends® in Information Retrieval, 1-154.

[Conneau et al. 2017] Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: "Supervised learning of universal sentence representations from natural language inference data"; arXiv preprint arXiv:1705.02364v5, 670-680.

[Devlin et al. 2019] Devlin, J., Chang, M., Lee, K., Toutanova, K.: "BERT: Pre-training of deep bidirectional transformers for language understanding"; Proc. North American chapter of the Association for Computational Linguistics: Human Language Technologies, 4171-4186.

[Dieng et al. 2020] Dieng, A., Ruiz, F., Blei, D.: "Topic modeling in embedding spaces"; Transactions of the Association for Computational Linguistics, 439-453.

[Doogan and Buntine 2021] Doogan, C., Buntine, W.: "Topic model or topic twaddle? Re-evaluating demantic interpretability measures"; Proc. North American chapter of the Association for Computational Linguistics, 3824-3848.

[Griffiths et al. 2007] Griffiths, T., Steyvers, M., Tenenbaum, J.: "Topics in semantic representation"; Psychological review, 211-244.

[Grootendorst 2022] Grootendorst, M.: "BERTopic: Neural topic modeling with a class-based tf-idf procedure"; arXiv preprint arXiv:2203.05794v1, 1-10.

[Harper and Konstan 2015] Harper, F., Konstan, J.: "The movielens datasets: History and context"; ACM Transactions on Interactive Intelligent Systems, 1-19.

[Howard and Ruder 2018] Howard, J., Ruder, S.: "Universal language model fine-tuning for text classification"; arXiv preprint arXiv:1801.06146v5, 1-12.

[Hoyle et al. 2021] 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, 2018-2033.

[Hoyle et al. 2025] Hoyle, A., Calvo-Bartolomé, L., Boyd-Graber, J., Resnik, P.: "ProxAnn: Use-oriented evaluations of topic models and document clustering"; Proc. 63^{rd} annual meeting of the Association for Computational Linguistics, 15872-15897.

[Isonuma et al. 2020] Isonuma, M., Mori, J., Bollegala, D., Sakata, I.: "Tree-structured neural topic model"; Proc. 58^{th} annual meeting of the Association for Computational Linguistics, 800-806.

[Kennedy et al. 2021] Kennedy, B., Atari, M., Davani, A., Hoover, J., Omrani, A., Graham, J., Dehghani, M.: "Moral concerns are differentially observable in language"; Cognition, 1-12.

[Kim et al. 2012] Kim, J., Kim, D., Kim, S., Oh, A.: "Modeling topic hierarchies with the recursive chinese restaurant process"; Proc. 21^{st} ACM Information and Knowledge Management, 783-792.

[Kiros et al. 2015] Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R., Urtasun, R., Torralba, A., Fidler, S.: "Skip-thought vectors"; Advances in Neural Information Processing Systems, 1-9.

[Kubat 2017] Kubat, M.: "An introduction to machine learning"; 2^{nd} ed.; Springer, 1-348.

[Kulkarni et al. 2021] Kulkarni, A., Hengle, A., Kulkarni, P., Marathe, M.: "Cluster analysis of online mental health discourse using topic-infused deep contextualized representations"; Proc. 12^{th} Workshop on Health Text Mining and Information Analysis, 83-93.

[Lee et al. 2020] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C., Kang, J.: "BioBERT: a pre-trained biomedical language representation model for biomedical text mining"; Bioinformatics, 1234-1240.

[Lewis et al. 2019] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension"; arXiv preprint arXiv:1910.13461v1, 1-10.

[Luong et al. 2015] Luong, M., Pham, H., Manning, C.: "Effective approaches to attention-based neural machine translation"; arXiv preprint arXiv:1508.04025v5, 1-11.

[Martin and Jurafsky 2019] Martin, J., Jurafsky, D.: "Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition"; 3^{rd} draft ed., 1-613; .

[Mikolov et al. 2013a] Mikolov, T., Chen, K., Corrado, G., Dean, J.: "Efficient Estimation of Word Representations in Vector Space"; arXiv preprint arXiv:1301.3781, 1-12.

[Mikolov et al. 2013b] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: "Distributed representations of words and phrases and their compositionality"; Advances in Neural Information Processing Systems, 1-9.

[Mikolov et al. 2013c] Mikolov, T., Yih, W., Zweig, G.: "Linguistic regularities in continuous space word representations"; Proc. North American chapter of the Association for Computational Linguistics: Human Language Technologies, 746-751.

[Mimno et al. 2011] Mimno, D., Wallach, H., Talley, E., Leenders, M., McCallum, A.: "Optimizing semantic coherence in topic models"; Proc. Empirical Methods in Natural Language Processing, 262-272.

[Nanda et al. 2023] Nanda, G., Jaiswal, A., Castellanos, H., Zhou, Y., Choi, A., Magana, A.: "Evaluating the coverage and depth of latent dirichlet allocation topic model in comparison with human coding of qualitative data: the case of education research"; Machine Learning and Knowledge Extraction, 473-490.

[Ng et al. 2001] Ng, A., Jordan, M., Weiss, Y.: "On spectral clustering: Analysis and an algorithm"; Advances in Neural Information Processing Systems, 1-8.

[Nothman et al. 2018] Nothman, J., Qin, H., Yurchak, R.: "Stop word lists in free open-source software packages"; Proc. Workshop for NLP Open Source Software, 7-12.

[Pagliardini et al. 2017] Pagliardini, M., Gupta, P., Jaggi, M.: "Unsupervised learning of sentence embeddings using compositional n-gram features"; arXiv preprint arXiv:1703.02507v3, 1-13.

[Pantoja et al. 2022] Pantoja, F., Carvalho, M., Santanchè, A.: "Harena Semantics: a framework to support semantic annotation in citizen science systems"; Proc. 5^{th} Biomedical Engineering Systems and Technologies (HEALTHINF), 336-343.

[Peinelt et al. 2020] Peinelt, N., Nguyen, D., Liakata, M.: "tBERT: Topic models and BERT joining forces for semantic similarity detection"; Proc. 58^{th} annual meeting of the Association for Computational Linguistics, 7047-7075.

[Pennington et al. 2014] Pennington, J., Socher, R., Manning, C.: "Glove: Global vectors for word representation"; Proc. Empirical Methods in Natural Language Processing, 1532-1543.

[Peters et al. 2018] Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: "Deep contextualized word representations"; arXiv preprint arXiv:1802.05365, 1-15.

[Qiu et al. 2024] Qiu, M., Yang, W., Wei, F., Chen, M.: "A topic modeling based on prompt learning"; Electronics, 1-16.

[Radford et al. 2018] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: "Improving language understanding with unsupervised learning"; Technical report, OpenAI, 1-12.

[Reimers and Gurevych 2019] Reimers, N., Gurevych, I.: "Sentence-BERT: Sentence embeddings using siamese BERT-networks"; arXiv preprint arXiv:1908.10084, 1-11.

[Röder et al. 2015] Röder, M., Both, A., Hinneburg, A.: "Exploring the space of topic coherence measures"; Proc. 8^{th} Conference on Web Search and Data Mining, 399-408.

[Ruder 2019] Ruder, S.: "Neural transfer learning for natural language processing"; PhD thesis, NUI Galway, 1-311.

[Simonetti et al. 2025] Simonetti, A., Albano, A., Plaia, A., Tumminello, M.: "Ranking coherence in topic models using statistically validated networks"; Journal of Information Science, 744-765.

[Šuster and Daelemans 2018] Šuster, S., Daelemans, W.: "CliCR: a dataset of clinical case reports for machine reading comprehension"; Proc. North American chapter of the Association for Computational Linguistics: Human Language Technologies, 1551-1563.

[Tenney et al. 2019] Tenney, I., Das, D., Pavlick, E.: "BERT Rediscovers the classical NLP pipeline"; arXiv preprint arXiv:1905.05950v2, 1-9.

[Thompson and Mimno 2020] Thompson, L., Mimno, D.: "Topic modeling with contextualized word representation clusters"; arXiv preprint arXiv:2010.12626v1, 1-12.

[Turian et al. 2010] Turian, J., Ratinov, L., Bengio, Y.: "Word representations: a simple and general method for semi-supervised learning"; Proc. 48^{th} Meeting of the Association for Computational Linguistics, 384-394.

[Vaswani et al. 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., Polosukhin, I.: "Attention is all you need"; Advances in Neural Information Processing Systems, 1-11.

[Wu et al. 2024] Wu, X., Pan, F., Nguyen, T., Feng, Y., Liu, C., Nguyen, C., Luu, A.: "On the affinity, rationality, and diversity of hierarchical topic modeling"; Proc. Conference on Artificial Intelligence, 19261-19269.

[Wu 2016] Wu, Y.: "Google's neural machine translation system: Bridging the gap between human and machine translation"; arXiv preprint arXiv:1609.08144v2, 1-23.

[Yan et al. 2022] Yan, H., Gui, L., He, Y.: "Hierarchical interpretation of neural text classification"; Computational Linguistics, 987-1020.

[Zhao et al. 2021] Zhao, H., Phung, D., Huynh, V., Jin, Y., Du, L., Buntine, W.: "Topic modelling meets deep neural networks: a survey"; arXiv preprint arXiv:2103.00498v1, 1-8.