Exploiting Contextual Embeddings in Hierarchical Topic Modeling and Investigating the Limits of the Current Evaluation Metrics

Felipe Viegas[1], Antonio Pereira[2], Washington Cunha[1], Celso França[1], Claudio Andrade[1], Elisa Tuler[2], Leonardo Rocha[2], and Marcos André Gonçalves[1]

[1]Federal University of Minas Gerais, Department of Computer Science
frviegas@dcc.ufmg.br, washingtoncunha@dcc.ufmg.br, celsofranca@dcc.ufmg.br, claudio.valiense@dcc.ufmg.br, mgoncalv@dcc.ufmg.br
[2]Federal University of São João Del Rei, Department of Computer Science
antoniopereira@aluno.ufsj.edu.br, etuler@ufsj.edu.br, lcrocha@ufsj.edu.br

We investigate two essential challenges in the context of hierarchical topic modeling (HTM)—(i) the impact of data representation and (ii) topic evaluation. The data representation directly influences the performance of the topic generation, and the impact of new representations such as contextual embeddings in this task has been under-investigated. Topic evaluation, responsible for driving the advances in the field, assesses the overall quality of the topic generation process. HTM studies exploit the exact topic modeling (TM) evaluation metrics as traditional TM to measure the quality of topics. One significant result of our work is demonstrating that the HTM's hierarchical nature demands novel ways of evaluating the quality of topics. As our main contribution, we propose two new topic quality metrics to assess the topical quality of the hierarchical structures. Uniqueness considers topic topological consistency, while the Semantic Hierarchical Structure (SHS) captures the semantic relatedness of the hierarchies. We also present an additional advance to the state-of-the-art by proposing the c-CluHTM. To the best of our knowledge, c-CluHTM is the first method that exploits contextual embeddings into NMF in HTM tasks. c-CluHTM enhances the topics' semantics while preserving the hierarchical structure. We perform an experimental evaluation, and our results demonstrate the superiority of our proposal with gains between 12% and 21%, regarding NPMI and Coherence over the best baselines. Regarding the newly proposed metrics, our results reveal that Uniqueness and SHS can capture relevant information about the structure of the hierarchical topics that traditional metrics cannot.

Action Editor: Miguel Ballesteros. Submission received: 20 November 2023; revised version received: 6 July 2024; accepted for publication: 3 September 2024.

https://doi.org/10.1162/coli_a_00543

1. Introduction

Topic modeling (TM) is one of the most prominent and useful approaches for extracting and organizing information from ever-increasing large amounts of data in an unsupervised fashion. These approaches aim at extracting and revealing underlying semantic topics from raw textual documents (e.g., product reviews, tweets), which can then be used by other applications, such as search engines and recommender systems, to help achieve high effectiveness in specific tasks (Porturas and Taylor 2021; Aziz et al. 2022).

A type of TM task that is very useful when there is a natural organization of the domain topics into different levels of semantic granularity is **Hierarchical Topic Modeling (HTM)** (Chauhan and Shah 2021; Churchill and Singh 2021). HTM is the task of automatically extracting latent topics (e.g., a concept or a theme) from a collection of textual documents, preserving the inherent hierarchical structure of the topics of the domain of interest. Such topics are usually defined as a probability distribution over a fixed word vocabulary. In HTM, topics are related to each other in a hierarchical fashion at different semantic granularity levels (more general or specific).

A recent method, called CluHTM (hereafter we will call the CluHTM standard solution f-CluHTM) (Viegas et al. 2020), has produced top-notch results in HTM tasks by taking advantage of semantic similarities obtained from distances between words within an embedding space (Mikolov et al. 2018; Shrivastava and Sharma 2021). In a nutshell, CluHTM exploits the nearest words of a given "pre-trained" static word embedding to generate "meta-words," or *CluWords*, able to expand and enhance the document representation in terms of syntactic and semantic information. CluWords also defines a heuristic based on stability for defining the number of topics and the "shape" of the hierarchical structure. Despite excellent results (Viegas et al. 2020), CluHTM's use of pre-trained **static** (global) embeddings leaves room for improvements, for instance, by exploiting more modern attention-based *contextual* word embeddings (e.g., based on BERT and its variants [Yang et al. 2019; Liu et al. 2019]), which can help to build richer word representations that capture contextual (local) information.

Indeed, contextual word embeddings aim to learn sequence-level semantics by considering the sequence of all words in the documents. Recent strategies, such as pPSO (Miles et al. 2022), have started exploring contextual embeddings to add semantic information to the data representation. However, extending CluHTM solution to exploit contextual word embeddings is not straightforward, since, different from static embeddings that have a 1-1 correspondence between a word and a vectorial word embedding representation, a word in a contextual representation may be associated with multiple embeddings in different contexts (or even in different layers of a deep transformer representation).

1.1 Research Objectives

In this context, with the advance of the state-of-the-art f-CluHTM, we propose c-CluHTM. This CluHTM variant implements a novel component in the CluHTM architecture that enables the application of contextual word embedding. It exploits BERT's hidden layers by exploiting several pooling strategies (e.g., average, maximum, concatenation combined with pooling) to build a single-word embedding representation for each word to be used in CluHTM's meta-word construction. The idea is to evaluate the quality of the topics built with these new word embedding representations in the CluHTM solution compared with the standard f-CluHTM solution. To the best of our

knowledge, this is the first solution that exploits contextual embeddings in the context of HTM.

Accordingly, one of the main objectives is to answer the following research question:

RQ1: *Does c-CluHTM build more informative topics when compared to f-CluHTM by exploiting contextual word embeddings?*

To answer **RQ1**, we position our analyses in the context of the state-of-the-art in HTM. In other words, in our experiments, besides directly comparing c-CluHTM with the state-of-the-art f-CluHTM, we also include four representative HTM baselines—HLDA, HPAM, hARTM, and BERTopic. We evaluate the HTM methods considering 12 datasets used in the original f-CluHTM work (Viegas et al. 2020) and evaluate topic quality using two main evaluation metrics: *Coherence* (Nikolenko, Koltcov, and Koltsova 2017) and *Normalized Pairwise point-wise Mutual Information* (NPMI). These metrics capture the quality of the interactions between words assigned to each topic. Consequently, they also measure the quality of each topic individually.

Our experimental results show that c-CluHTM achieves consistent results with higher scores than the baselines for all evaluated metrics, corroborating our hypothesis that contextual embeddings benefit the HTM task. Indeed, all the c-CluHTM variants, using the different pooling strategies, reached NPMI results of 0.93 or higher in all datasets.

These very high effectiveness results, close to the maximum value of 1, raise an important question regarding the current HTM evaluation methodologies—*Are we approaching the limits of what the current evaluation metrics can assess regarding topical quality for HTM?* Traditional metrics capture the quality of topics by strictly evaluating the built topics, ignoring the hierarchical structure. Indeed, in a brief empirical evaluation of the topics built with c-CluHTM, we observed that the evaluated metrics could not capture some important idiosyncratic aspects of the HTM task related to the existence of near-duplicate topics and the repetition of words in several topics over a hierarchical structure.

These aspects are here defined as the *consistency of the topic generation*. Indeed, consistency is strictly related to redundancy, usually manifested as (near-)duplicated topics composed of (almost) the exact words. Duplicity implies a lack of consistency in topic generation. Such issues may also promote artificial biases in the evaluation. For instance, duplicated information may improperly boost (in the case of a "good" duplicated topic) or decrease (in a non-informative topic) the final result of a topic evaluation for an HTM method. None of the traditional evaluation metrics can capture these issues simply because they were not designed for doing so.

To deal with the limitations of the current evaluation metrics, we design two new metrics that capture different perspectives of the HTM task. Given two (consecutive) levels of the hierarchy structure generated by an HTM method, the topics can be assessed as follows. A main (focus) topic 1 (upper level) is compared to the subtopics (e.g., 1.1) generated from it (lower level), using *intra-group distances*. On a different perspective, a main topic 1 can be compared to the subtopics generated based on another topic 2 in the same hierarchical level of 1 (e.g., 2.1, 2.2, etc.), considering *inter-group distances*. The intuition here is that *intra-group distances* must be small while *inter-group distances* must be larger than the *intra-group distances*.[1]

1 Figure 6 in Section 4 provides a pictorial notion of the concepts expressed above.

As our **main** contribution, we raise the second research question we aim to answer:

RQ2: *Are the concepts of intra- and inter-topic distance helpful in measuring the quality of the topic hierarchies built by HTM methods?*

We adopted these two concepts of *intra-* and *inter-topic* distances to analyze topic information at a hierarchical level. *Intra-topic distance* measures the correlation between topics in the same hierarchy, that is, between topics in different levels of the same topology.[2] *Inter-topic distance* measures the correlation between topics from different topologies. In general, what is expected is that topics in the same hierarchical structure should deal with more related subjects than those in adjacent hierarchical structures.

Based on these two new concepts of *intra-group* and *inter-group distances*, as our main contribution, we design two new topic quality metrics to assess topical quality, considering two aspects: (i) *topic topological consistency (or redundancy)* and (ii) *semantic hierarchical structure*. The Uniqueness metric assesses the consistency of topic generation given two consecutive levels of the hierarchy built by an HTM method. This metric can automatically measure the level of redundancy of topic generation. The second metric, Semantic Hierarchical Structure (SHS), uses cosine distance considering a semantic representation of the topics based on word embedding representations to calculate the semantic relatedness of topics of the same topology (intra-topic) and between topics of distinct topologies (inter-topic), respectively.

Our experimental evaluation of the proposed metrics using the same experimental setting as before shows that the proposed metrics can capture distinct properties of the topics built by the HTM methods. Our results show that CluHTM variants (c-CluHTM and f-CluHTM) presented the best results when compared to the baselines regarding the Uniqueness and SHS, showing consistency in terms of the **RQ2** hypothesis that *inter-group distances* are longer than *intra-group distances*. Finally, our experimental results reveal that there is significant room for new HTM proposals that are more robust from a perspective of consistency of the topological topical structure.

1.2 Main Contributions

We summarize the main contributions of this research article as follows:

- We propose a hierarchical topic modeling strategy that exploits contextual embedding information called c-CluHTM. The new approach implements a novel component in the CluHTM architecture that enables the use of contextual word embeddings;
- The proposal of **two** new topic evaluation metrics to measure the quality of topics based on the aspects (i) *topic topological consistency* and (ii) *semantic hierarchical structure*. These two proposed metrics complement the traditional quality metrics in the context of hierarchical topic modeling;
- We offer an extensive experimental evaluation considering 12 datasets and main HTM methods. We contrast the topic evaluation regarding

2 The topological structure is defined by a tree structure in which the topics are built. The depth walk in this tree structure can be considered a topology of topics.

traditional quality metrics (NPMI and Coherence) and the proposed metrics.

- We report on a human evaluation experiment to contrast the new metric with human judgment.

1.3 Article Organization

This article is organized as follows. Section 2 covers related work. Section 3 describes the family of CluHTM solutions for the HTM task, including the previous version with static embeddings and our new proposal, c-CluHTM, that exploits attention-based contextual embeddings. Section 4 provides the details for our new proposed HTM evaluation metrics. Section 5 includes the experimental setup, and Section 6 presents the results. Section 7 provides discussions. Section 8 concludes the article.

2. Related Work

2.1 Hierarchical Topic Modeling

hierarchical topic modeling (HTM) is a sub-area of topic modeling where topics are built in a tree structure. HTM methods can also be grouped into supervised and unsupervised. Considering supervised strategies, we highlight some extensions to the traditional latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003) used in TM. Mcauliffe and Blei (2007) proposed SLDA, a supervised LDA extension for traditional TM that allows one to connect each document to a regression variable to find latent topics. Based on SLDA, hierarchically supervised LDA (HSLDA) (Perotte et al. 2011) incorporates the data hierarchy with multiple labels and pre-labels into a single model. SNLDA (Resnik et al. 2015) is based on SLDA, implementing a generative probabilistic strategy in which topics are sampled from a probability distribution. SNLDA extends SLDA by assuming that topics are organized in a tree structure. These extended versions better predict the response variables for future unlabeled documents (Yu et al. 2022; Burkhardt and Kramer 2019).

Considering the unsupervised HTM strategies, HPAM (Han, Han, and Li 2019) is an extension of the TM technique known as Pachinko Allocation (PAM) (Vayansky and Kumar 2020). In PAM, documents are a mixture of distributions on a set of particular topics, using a directed acyclic plot to represent the topic co-occurrences. Each node of this graph represents a Dirichlet distribution. There is only one node at the highest level of PAM, whereas the lowest levels represent a distribution among the next higher-level nodes. In HPAM, each node is associated with a distribution in the vocabulary of the document.

Liu et al. (2020) proposed HLDA, also an extension of LDA, to be considered a representative baseline in the context of HTM. In HLDA, Nested Chinese Restaurant Process (NCRP) is used to generate a hierarchical tree in addition to using the Dirichlet distribution. More recently, Xu et al. (2018) proposed KHTM, a method based on HLDA, and, as such, models a generative process whose parameter estimation strategy is based on *Gibbs* sampling.

Liu et al. (2018) introduced HSOC, which uses non-probabilistic matrix factorization (NMF) to solve HTM tasks. To alleviate the main NMF drawbacks in the HTM configuration, HSOC relies on three optimization constraints to properly control matrix factoring operations properly when discovering the topic's hierarchical structure. Such

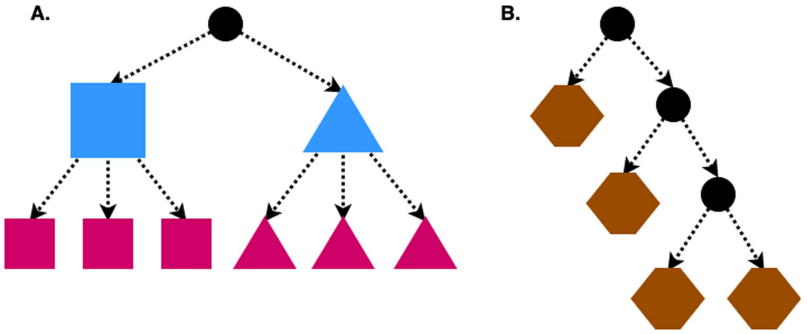

Figure 1
A. Example hierarchical structure built by CluHTM approaches, HLDA, HPAM, and hARTM.
B. Hierarchical structure built by BERTopic.

constraints are *global independence*, *local independence*, and *information consistency*, and allow HSOC to create hierarchical topics that somehow preserve topic coherence and hierarchical semantic structures.

BERTopic (Grootendorst 2022) is a hybrid solution that exploits BERT embeddings to represent the documents and exploits a hierarchical clustering approach to build hierarchical topics. In more detail, the method is composed of five steps: (i) Embedded documents: This step exploits the information provided by a transformer approach (i.e., BERT) to get the embedding vector (i.e., the CLS embedding) of the documents; (ii) Dimensionality reduction: This step exploits PCA or UMAP to reduce the dimension of the document embedding, (iii) Cluster Documents: This step exploits the HDBSCAN method to build the clusters of documents; (iv) Bag-of-Words (BoW): This applies a BoW representation in a cluster-level to have a better representation of the words that occur in the clusters; and (v) Topic Representation: This exploits the c-TFIDF approach over the BoW representation of the previous step to get the topic representation based on words that (co-)occur in the same clusters. This method is capable of building topics into hierarchies. However, the topological structure is different from other HTM methods. The levels of BERTopic hierarchy have only two topics (like a binary tree), as illustrated in Figure 1B, unlike the other methods adopted as baselines, that produce hierarchies such as those shown in Figure 1A. Since this method exploits BERT embeddings like ours, we use it as a baseline.

hARTM (Vorontsov et al. 2015) is a non-Bayesian method hierarchical version of Additive Regularization of Topic Models (ARTM). This method exploits regularization to mitigate problems posed by the LDA-based methods. Bayesian generative models, like LDA, assume that topic distributions over words and document distributions over topics are generated from prior Dirichlet distributions. This conflicts with two natural practical realities due to sparsity: (i) topics with zero probabilities in a document and (ii) words with zero probabilities in a topic. The authors mention that regularization reduces a potentially infinite set of solutions and helps to select a better one. The hierarchical version has a restriction of building only two levels of hierarchy per topic. We use this method as a baseline, adopting the implementation provided by BigARTM.[3]

3 `https://github.com/bigartm/bigartm`.

Finally, f-CluHTM (Viegas et al. 2020) is a recent NMF-based strategy that adopts the representation of CluWords (Viegas et al. 2019) together with an NMF variant specially designed for modeling topics on a document hierarchy. This strategy uses a stability measure to calculate the optimal number of topics at each hierarchy level to build a hierarchy. In Viegas et al. (2020), the solution was compared with most of the above methods, presenting results far superior to all others, being considered today's state-of-the-art hierarchical topic modeling. In that comparative analysis, HLDA was the second-best method.

Here, we propose c-CluHTM, a method that exploits contextual word embeddings from BERT. The proposed method includes a transformation step that converts contextual word embeddings into static word embeddings. This step enables us to use any contextual word embeddings in the solution, particularly within an NMF-based strategy. We exploit three types of transformations: Max, Average, and Concatenation. The concepts of contextual word embeddings and static word embeddings are further described in Section 2.3.

2.2 Topic Evaluation Metrics

The two main topic evaluation metrics used in the literature assess topic quality according to the occurrences of the top chosen words in each topic t, without considering the topic's correlations (Nikolenko 2016). The metrics are:

Coherence: captures the ease of interpretation according to word co-occurrences, as defined in Equation (1), where W_t is the set of top p words selected to represent a topic t, d represents a document $\in \mathbb{D}$.

$$c_{\text{tf-idf}}(t, W_t) = \sum_{w_1, w_2 \in W_t} log \frac{\sum_{d:w_1,w_2 \in d} \text{tf-idf}(w_1, d) \times \text{tf-idf}(w_2, d)}{\sum_{d:w_1 \in d} \text{tf-idf}(w_1, d)} \tag{1}$$

tf-idf is the increasing frequency according to Equation (2):

$$\text{tf-idf}(w, d) = \left(\frac{1}{2} + \frac{f(w, d)}{max_{w' \in d} f(w', d)}\right) \times log\left(\frac{|D|}{|\{d \in D : w \in d\}|}\right) \tag{2}$$

and $f(w, d)$ is the number of occurrences of word w in document d.

Normalized Pairwise point-wise Mutual Information (NPMI): measures how much information a word w_i "gains" given the occurrence of another word w_j, taking into account

word dependencies. For a given ordered set of the most p important words W_t of a topic t, the NPMI metric is calculated as:

$$\text{NPMI}_t = \sum_{i<j} \frac{log \frac{P(w_i,w_j)}{P(w_i)P(w_j)}}{-log\ P(w_i,w_j)} \tag{3}$$

These metrics assess topic quality produced by TM methods exploring different aspects. Thus, these metrics complement each other in evaluating the effectiveness of TM methods.

In this work, we propose evaluation metrics that complement the above-mentioned ones. While traditional metrics are focused on assessing the "agreement" among the words that define a topic, our proposed metrics can syntactically and semantically assess the divergence of topics, providing an assessment in terms of aspects not yet considered in the literature, such as: (i) *topic topological consistency (or redundancy)* and (ii) *semantic hierarchical structure*.

2.3 Word Embeddings Representations

Vector space models, also known as word embeddings, represent (embed) words in a continuous vector space where semantically similar words should be mapped to nearby points. Word embeddings are built by a process similar to an auto-encoder, encoding each word in a vector, but rather than training against the input words through reconstruction, as a restricted Boltzmann machine does, Word2Vec trains words against other words that neighbor them in the input corpus.[4] All word embedding methods depend on the Distributional Hypothesis (Dieng, Ruiz, and Blei 2020), which states that words that appear in the same contexts share semantic meanings. Approaches to adopting this principle may be divided into two categories: methods based on counting (latent semantic analysis) and predictive methods (Word2Vec).

fastText (Mikolov et al. 2018), on the other hand, learns vectors for the sub-words (i.e., character n-grams) found within each word, as well as the complete word. At each training step in fastText, the mean of the target word and sub-word vectors are used for training. In Mikolov et al. (2018), the authors claim that fastText generates better word embeddings for rare words. Indeed, using character embedding for downstream tasks has been shown to boost the performance of those tasks compared to using Word2Vec. Representations produced by both Word2Vec and fastText are considered as *static word embeddings* since there is only one embedding for each word.

BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al. 2019) is a method that exploits pre-trained language representations. It uses a transformer, an attention mechanism that learns the correlations between words in a text. A transformer is built with two separate mechanisms, an encoder that reads text input and a decoder that produces a prediction for the task. The BERT differential lies in its bidirectional approach. Instead of using only the previous or subsequent words to construct the representation of a term, BERT reads the entire sequence of words at once. Its methodology comprises two phases. The first consists of replacing 15% of the words in each sequence with a MASK token, and then the model tries to predict the value of the masked words using the surrounding unmasked ones. In the second phase, the model

4 Sentences used to train the model.

receives pairs of sentences as input and learns to predict whether the second sentence is subsequent in the original document. In the end, we have very semantically rich word representations (a.k.a. contextual word embeddings) that take into account the context for each word.

Recent TM strategies have started to exploit contextual embeddings. pPSO (Miles et al. 2022) is a recent method that adds semantic information to the data representation. It is a modified version of Particle Swarm Optimization (PSO), which tracks particle fitness on a dimension-by-dimension basis and is then applied to contextual embeddings to create topics. ETM (Dieng, Ruiz, and Blei 2020) is a generative document model that marries traditional topic models with contextual embeddings. Despite motivating the use of contextual embedding representations for HTM tasks, pPSO and ETM cannot be straightforwardly applied to the hierarchical context without significant extensions. Some recent work has also shown that applying pooling strategies to transform Contextual embeddings into static embeddings can lead to promising results in other domains, such as document classification and comparison among embeddings models (Bommasani, Davis, and Cardie 2020; Zhou and Bloem 2021; de Andrade et al. 2023).

3. CluHTM Solutions

This section briefly presents the CluHTM solution, which combines the CluWords representation with a non-probabilistic factorization method. The CluHTM solution is the basis for our hierarchical solutions and some of the baselines.

3.1 Background: CluWords Concept

CluWords (Viegas et al. 2019) builds document representations that take advantage of word embedding models to capture semantic relationships between words. However, embedding models may introduce semantic noise when words with different meanings have high similarity in the vector space. To deal with the problem, Cluwords has mechanisms to filter out potential noise (i.e., semantic noise, irrelevant words) and properly weight words according to the application scenario. In more detail, to build a document representation, CluWords applies three generic steps:

Word Clustering This step explores distance-based metrics between word embedding vectors[5] to group semantically similar words aiming at enriching documents with semantic information. The Word Clustering step requires a static word vector space that supports that language of the dataset, namely, a single vector representation for each word in the vocabulary $\mathbb{V}$.[6] For the sake of efficiency, we adopt HNSW[7] (Malkov and Yashunin 2018), a fast implementation of the approximate nearest neighbor search. It is important to stress that this step is deterministic as there is no random sub-step (e.g., seed initialization) on it. Thus, a nearest neighbor search is performed to create the semantic matrix C, a quadratic matrix

5 Note that we are not necessarily limited to embedding models. However, this is the main current approach in the literature to capture relationships between words. It can be easily modified in the future.

6 The vocabulary $\mathbb{V}$ can be represented as uni-gram or n-grams, this will depend of the preprocessing applied in the dataset.

7 `https://github.com/nmslib/hnswlib`.

of size $|\mathbb{V}| \times |\mathbb{V}|$, where $|\mathbb{V}|$ is the size of the vocabulary, where each cell $C_{i,j}$ stores the cosine similarity between the word vectors of words w_i and w_j. The limitation of receiving only static embedding representations has to do with the way that the semantic relationship is captured by this 2-D semantic matrix C. The next steps were designed to mitigate this limitation.

Filtering consists of applying strategies capable of filtering noise due to issues in the previous word clustering step or due to inconsistent words (from an application perspective) existent in a semantic neighborhood. The semantic neighborhood can be represented by a matrix C as defined below. An α parameter controls the semantic information added in the matrix C, captured by the cosine similarity. Values close to 1.0 are more conservative and will add less, but cleaner, information in the C, while values close to 0.0 will add more information. The α value must be chosen depending on the pair-wise cosine similarity distribution of the embedding vector space used to build the semantic neighborhood.

$$C_{t',t} = \begin{cases} \omega(u_{t'}, u_t) & \text{if } \omega(u_{t'}, u_t) \geq \alpha \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Weighting combines the (word) semantic information built in the first step with document information. This step explores strategies that weigh the semantic representation contained in CluWords and the corresponding document representation after CluWords processing. Currently, the document representation used in the weighting step is based on a Bag-of-Words (BoW) representation. In this work, we exploit the TFIDF representation to weight the importance of words for the document. The final CluWords representation is defined as follows. Equation (5) computes the TFIDF of the textual representation, while Equation (6) is the dot product that computes the final CluWords representation. Note that the CluWords representation is a sparse document representation because of the BoW representation aside the semantic matrix C. So, this is the main goal of exploiting a BoW representation in this step.

$$TFIDF_{t,d} = TF_{t,d} \cdot log(\frac{|\mathcal{D}|}{n_t}) \tag{5}$$

$$CW_{d,t} = \overrightarrow{TFIDF_d} \times \overrightarrow{C_{,t}} \tag{6}$$

Figure 2 summarizes the process of transforming textual documents into CluWords (cluster of words) representations. The word clustering and filtering steps exploit the nearest neighborhood approach to build the semantic information about the document set, combined with word (embedding) vectors. The weighting step combines semantic and statistical information to create the final CluWords representation for the dataset.

3.2 Background: Stability Measure

The Stability measure (Greene, O'Callaghan, and Cunningham 2014) is used in our proposed strategy to enable the CluHTM solution to automatically set the number of topics over the hierarchy of topics.

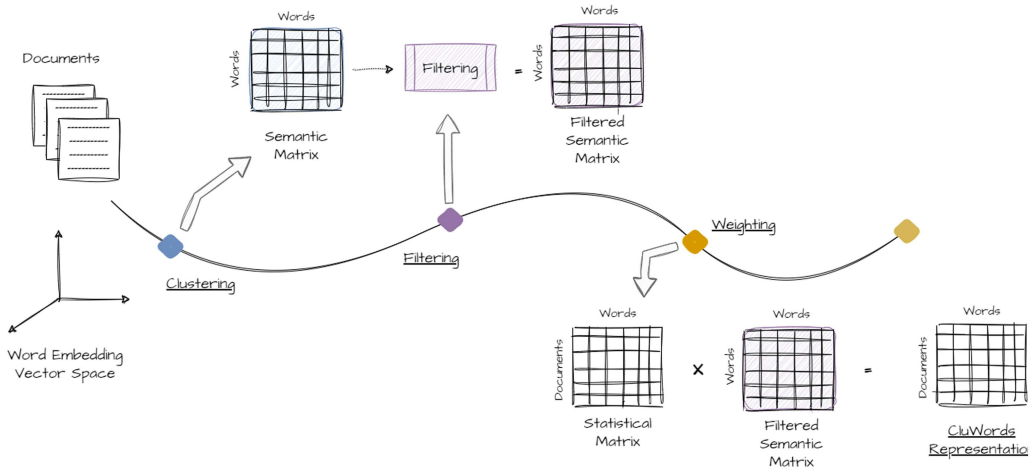

Figure 2
The steps for building the CluWords representation.

The intuition behind this strategy is to select the proper number given a predefined range of number of topics $[\mathcal{K}_{min}, \mathcal{K}_{max}]$. The proper number of topics is selected by multiple topic modeling runs that contrast random sampling runs S and the main topic modeling runs D. In each topic modeling run, the Stability approach checks the agreement of the topics built. The agreement is defined as $agree(\mathcal{W}_x, \mathcal{W}_y) = \frac{1}{p}\sum_{i=1}^{p} AJ(w_{xi}, \rho(w_{xi}))$, where W_x and W_y are the set of the top p words that represents the topics x and y, respectively. $AJ(\cdot)$ is the average Jaccard coefficient used to compare the similarity among the topic words, w and $\rho(\cdot)$ is the optimal permutation of the words in $\mathcal{W}_{\mathcal{S}_i}$. The p value defines the number of words that compose the topics; we used $p = 10$, the same number used in Viegas et al. (2020).

3.3 Background: f-CluHTM Solution

The original CluHTM method, hereafter called f-CluHTM, is an iterative method able to automatically define the "ideal" number of topics in each level of the hierarchy, given a pre-defined range $[\mathcal{K}_{min}, \mathcal{K}_{max}]$. f-CluHTM explores Cluwords and NMF (Meng et al. 2018), one of the main non-probabilistic strategies. The Stability method (Greene, O'Callaghan, and Cunningham 2014) is used to select NMF k parameters (i.e., a number of topics). Stability measures whether running multiple random samplings for a topic modeling strategy results in Stability in terms of p top words extracted from the topics.

f-CluHTM has five inputs (Algorithm 1): (i) $\mathcal{D}_{max}$ corresponds to the depth down to which we want to extract the hierarchical structure; (ii) $\mathcal{K}_{min}$ and $\mathcal{K}_{max}$ control the range of topics; such a range will be used in all levels of the hierarchy; (iii) $\mathcal{T}$ is the input text data; and (iv) $\mathcal{W}$ is the word embedding vector space used in the CluWords generation. The output is the hierarchical structure $\mathcal{H}$ of p top words for each topic. The hierarchical structure $\mathcal{H}$ contains the set of topologies τ exploited in Section 4.

The method starts by obtaining the root topic (lines 2–3 of Algorithm 1), which is composed of all documents in $\mathcal{T}$. Because the method is iterative, each iteration is controlled by a queue schema to build a hierarchical structure. At each iteration (line 3), the algorithm produces the CluWords representation for the documents $\in \mathcal{T}'$ (line 5), chooses the number of topics, exploiting the Stability measure (line 6), and runs

Algorithm 1: f-CluHTM

Input: $\mathcal{D}_{max}$ - Hierarchy Depth;
$\mathcal{K}_{min}$ - Number of minimum topics;
$\mathcal{K}_{max}$ - Number of maximum topics;
$\mathcal{T}$ - Term-frequency representation;
$\mathcal{W}$ - Word embedding vectors $\in \mathcal{T}$;
Output: $\mathcal{H}$ - Hierarchical Structure;
parent - Parent dependency

```
1  parent ← −1;
2  queue.push(0, T);
3  while queue ≠ ∅ do
4      depth, T' ← queue.pop();
5      Clu ← GenerateCluwords(T', W);
6      K ← Stability(K_min, K_max, Clu)
7      O ← NMF(Clu, K)
8      topics ← ExtractTopics(O)
9      foreach topic ∈ topics do
10         parent ← parent ∪ topic;
11         H ← H ∪ topic;
12         if depth + 1 ≤ D_max then
13             T' ← ExtractDocs(topic);
14             queue.push(depth + 1, T')
15 return H, parent
```

the NMF method (line 7) to extract the p words for each topic in $\mathcal{O}$ (line 8). Then, in the loop of line 9, each topic and its documents are stored in the queue.

Summarizing, f-CluHTM exploits *global* semantic information (captured by CluWords) with *local* factorization, limited by a stability criterion that defines the "shape" of the hierarchical structure. Though simple, combining these ideas is extremely powerful for solving the HTM task, as we see in Viegas et al. (2020).

3.4 The c-CluHTM Solution

c-CluHTM extends f-CluHTM by receiving contextual embedding representations (i.e., more than one representation per word) as input and transforming these contextual representations into static ones. This transformation is required as CluWords currently does not handle multiple representations for the same word—in the Clustering step, each word must have a single representation.

We exploit three pooling operations to transform BERT's contextual embeddings into a static one. We use the same notation for all three operations, described as follows. Let $\mathbb{D} = \{d_1, d_2, \cdots, d_{|\mathbb{D}|}\}$ be defined as the set of documents, $|\mathbb{D}|$ being the collection size (number of documents). Let $\mathbb{V} = \{w_1, w_2, \cdots, w_{|\mathbb{V}|}\}$ be the vocabulary of words in the collection, $|\mathbb{V}|$ being the vocabulary size. Each word $w_i \in \mathbb{V}$ has an embedding representation $\vec{w}_j \in \mathbb{K}_{i,x}$, where $\mathbb{K}_{i,x}$ is the set of each occurrence of word w_i in $\mathbb{D}$, and $x = \{1, 2, 3, \cdots, 12\}$ corresponds to a BERT embedding layer.

Max-pooling (Equation (7)) retrieves the maximum embedding $\vec{w}_j \in \mathbb{K}_{i,12}$ for the word $w_i \in \mathbb{V}$. In other words, the max-pooling operation takes the maximum value of each embedding dimension. Average pooling (Equation (8)) is the average operation of the embedding representations of the word w_i considering BERT's embedding layer 12 ($\mathbb{K}_{i,12}$). And finally, the Concatenation pooling is the average considering concatenated embedding representation $\overrightarrow{cw}_j \in \mathbb{K}_{i,concat}$, where $\mathbb{K}_{i,concat}$ represents the concatenation of all embedding representation occurrences of word w_i, regarding BERT's embedding layers 9, 10, 11, and, 12 ($\mathbb{K}_{i,9}||\mathbb{K}_{i,10}||\mathbb{K}_{i,11}||\mathbb{K}_{i,12}$). Then, similar to Average pooling, it computes an average embedding representation ($\overrightarrow{\mu wC}_i$) for the word w_i.

$$\overrightarrow{wM}_i = Max(\vec{w}_j), \text{ where } \vec{w}_i \in \mathbb{K}_{i,12} \tag{7}$$

$$\overrightarrow{\mu w}_i = \frac{\sum_j^{\mathbb{K}_{i,12}} \vec{w}_j}{|\mathbb{K}_{i,12}|} \tag{8}$$

$$\overrightarrow{\mu wC}_i = \frac{\sum_j^{\mathbb{K}_{i,concat}} \overrightarrow{cw}_j}{|\mathbb{K}_{i,concat}|} \tag{9}$$

In summary, the main difference between f-CluHTM and c-CluHTM is the use of different word embeddings. f-CluHTM exploits the pre-trained fastText, whereas the c-CluHTM exploits the BERT's contextual embeddings pooled in a single representation using three different strategies—Max, Average, and, Concatenation of word vectors.

4. Proposed Topic Evaluation Metrics

We start by introducing the concept of Topology (τ) in the context of HTM, the basis of the explanation for the new metrics. We then present the proposed metrics for evaluating topics in HTM.

The proposed evaluation metrics require two *consecutive* levels of a (hierarchical) topology to compute the quality of the topics. Figure 3 illustrates two levels of a topic

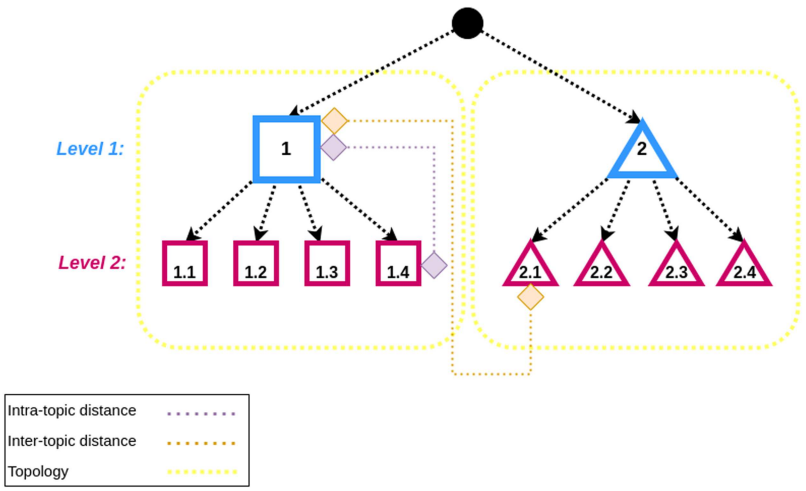

Figure 3
Example of topic's topology built by an HTM method.

hierarchy built by an HTM method m. In the first level of the hierarchy, method m generates two topics, the blue square and the blue triangle. Each topic is represented by the top p words, where p is defined as an input parameter of method m. In the second level of the hierarchical structure, m builds eight topics, four topics are derived from the blue square, represented as the pink squares, while the last four topics are derived from the blue triangle, illustrated as the four pink triangles. Note that each of these topics is also represented as top p words, which are used to measure the quality of the topics.

Topology τ_i is defined by the set of topics $\mathbb{T}_i$ originating from topic t_1, i, including itself. While the topology τ_j is defined by the set of topics $\mathbb{T}_j$ originating from topic t_1, j (i.e., local perspective). In Figure 3, the topologies marked in yellow illustrate topologies τ_i and τ_j of topics $t_{1,i}$ (blue square) and $t_{1,j}$ (blue triangle), respectively. The topics illustrated as squares belong to topology τ_i, since the pink squares were derived from the blue square, while the triangles belong to topology τ_j, as the blue triangle is the topic at a hierarchy level above the pink triangles. It is important to stress that the topologies must have more than one (sub-)topic in the second level and lower levels. So, hierarchical structures with only one topic at each level are not suitable for this type of evaluation.

Now, we can define our proposed quality evaluation metrics, which exploit concepts from clustering algorithms. Topics are seen as clusters composed of the top t words. These word clusters (topics) can be evaluated based on two perspectives: (i) *intra-group distances*, which measures the distance between objects of the same cluster, and (ii) *inter-group distances*, which measures the distance of objects from distinct clusters. We adapt these two concepts of *intra-* and *inter-topic* distances to analyze topic information in a hierarchy. *Intra-topic distances* measure correlations between topics in the same hierarchy, that is, between topics in different levels of the same topology, as described in Section 4.1. *Inter-topic distances* measure correlations between topics from different topologies (described in Section 4.2).

Based on these two new concepts, we design two new topic quality metrics to assess topical quality, considering two aspects: (i) *topic uniqueness* and (ii) *semantic hierarchical structure*. Before presenting the proposed metrics, let's introduce some notation that we will use throughout this section. Let $t_{i,j} \in \mathbb{T}_j$ be a latent topic built by the HTM method m, and $t_{i,j}$ be composed by the top κ words that best describe the topic $t_{i,j}$. $\mathbb{T}_j = \left\{t_1, t_2, t_3, \cdots, t_{|\mathbb{T}_j|}\right\}$ is the set of latent topics of the topology $\tau_j \in \tau$. The symbol $\tau = \{\tau_1, \tau_2, \cdots \tau_N\}$ defines the topologies created by the HTM method m. The two proposed quality metrics are described as follows:

Uniqueness: This metric evaluates whether HTM strategies can produce distinct topics without repeating words. Uniqueness assesses the consistency in terms of (lack of) redundancy of words among the topics regarding the information built by the strategy. HTM strategies might inadvertently produce redundant topics that do not offer new insights or differentiate sufficiently in their vocabulary. Take for instance the topic t_1 :{*data, mining, statistics, learning, algorithms*} and the topic t_2 :{*data, science, analysis, learning, algorithms*}. As we can see, both topics t_1 and t_2 are quite redundant since they share the words *data*, *learning*, and *algorithms*. Uniqueness assesses the consistency and the lack of redundancy of words among topics, which is vital for the specialization and effectiveness of topic hierarchies. This metric is a straightforward computation of compound word singularity (uniqueness) in topics. *Uniqueness* is defined by the function $unique(t_{i,*}, t_{j,*})$, which returns the gross number of unique words that compose both topics $t_{i,*}$ and $t_{j,*}$,

where $*$ represent any topic in the topologies τ. The gross number of unique words is divided by the total number of words regarding both topics ($t_{i,*}$ and $t_{j,*}$). Thus, Uniqueness captures the ratio of singularity in terms of words chosen to represent topics. The higher the value of this ratio, the more singular the topics are.

$$unique(t_{i,*}, t_{j,*}) = \frac{\left(t_{i,*} \cup t_{j,*}\right) - \left(t_{i,*} \cap t_{j,*}\right)}{\left(t_{i,*} \cup t_{j,*}\right)} \tag{10}$$

The Uniqueness metric is defined in Equation (11), where $\omega(\cdot)$ computes the word count.

$$uniqueness(t_{i,*}, t_{j,*}) = \frac{unique(t_{i,*}, t_{j,*})}{\omega(t_{i,*}) + \omega(t_{j,*})} \tag{11}$$

SHS: This metric captures the semantic information contained in each topic. It uses word vectors (embeddings) that compose the topics to evaluate topic quality in terms of semantic similarity. More specifically, SHS (Equation (12)) is a variant of the cosine distance among the embedding representation of topics $t_{i,*}$ and $t_{j,*}$. The idea is to assess the diversity and consistency of the topological hierarchy τ created by HTM strategies. From this semantic perspective, the closer the word vectors are to the topics in the same topology, the more consistent the hierarchy is.

$$SHS(t_{i,*}, t_{j,*}) = 1.0 - \frac{\sqrt{\sum \zeta(t_{i,*}) * \zeta(t_{j,*})}}{\sqrt{\sum \zeta(t_{i,*})^2} * \sqrt{\sum \zeta(t_{j,*})^2}} \tag{12}$$

The embedding representation of the topic $t_{i,*}$ is measured by the average pairwise cosine distance of their top κ words. $\zeta(t)$ illustrates the mean embedding representation of the topic $t_{i,*}$, where w_i is a word that represents the topic $t_{i,*}$, and $\overrightarrow{w_i}$ is its respective word embedding vector. The notation $|t_{i,*}|$ represents the number of words that represent the respective topic.

$$\zeta(t_{i,*}) = \frac{\sum_{w_i \in t_{i,*}} \overrightarrow{w_i}}{|t_{i,*}|} \tag{13}$$

The two proposed metrics are defined in the two mentioned contexts: by considering (i) intra- and (ii) inter-topic distances, as described next.

4.1 Metrics Considering Intra-topic Distances

The *intra-topic distances* measure correlations among topics $\mathbb{T}_z$ within the same topology $\tau_z \in \tau$. Equation (14) presents the intra-topic distance measured by the average scores of the $\sigma(\tau_i, \tau_j)$ between topics $\mathbb{T}_z \in \tau_z$. Note that Equation (14) is in fact a template to build the two proposed quality metrics in terms of intra-topic distances. The function $\sigma(\cdot)$ is a template that can be replaced for Uniqueness (Equation (15)) or SHS (Equation (16)).

$$intra = \frac{\sum_{t_i \in \mathbb{T}_z} \sum_{t_j \in \mathbb{T}_z, t_i \neq t_j} \sigma\left(t_{i,z}, t_{j,z}\right)}{|\mathbb{T}_z|} \tag{14}$$

where $|\tau|$ corresponds to the total number of topologies.

Uniqueness: In this metric, the distance function in Equation (14) is replaced by the $uniqueness(t_i, t_j)$ function. Uniqueness in terms of Intra-topic distances (DU_{intra}) is defined as:

$$DU_{intra} = \frac{\sum_{t_i \in \mathbb{T}_z} \sum_{t_j \in \mathbb{T}_z, t_i \neq t_j} uniqueness(t_i, t_j)}{|\mathbb{T}_z|} \tag{15}$$

where t_i e t_j belongs to the same topology and same level of the hierarchy.

SHS: To capture semantics, we use the SHS (Equation (12)) of the embedding centroids measured by Equation (13). Thus, the SHS intra-topic distance (DC_{intra}) is defined as:

$$DC_{intra}(\tau_k) = \frac{\sum_{t_i \in \mathbb{T}_z} \sum_{t_j \in \mathbb{T}_z, t_i \neq t_j} SHS(t_i, t_j)}{|\mathbb{T}_z|} \tag{16}$$

4.2 Metrics Considering Inter-topic Distances

Inter-topic distances measure correlations among topics from different topologies. The goal is to assess how far two different topologies τ_z and τ_y are. Equation (17) refers to the *inter-topic distance* measured by the distance among the topics of a topology $\mathbb{T}_z \in \tau_z$ to distinct topologies $\mathbb{T}_y \in \tau_z$. Similar to Equation (14), this formula measures the average of scores. In addition, the function $\sigma(\cdot)$ is a template that can be replaced for Uniqueness (Equation (18)) or SHS (Equation (19)) metrics. We present the instantiation of two metrics that consider inter-topic distances.

$$inter = \frac{\sum_{t_i \in \mathbb{T}_z} \sum_{t_j \in \mathbb{T}_y} \sigma(t_i, t_j)}{|\mathbb{T}_z| * |\mathbb{T}_y|} \tag{17}$$

where $|\mathbb{T}_z| * |\mathbb{T}_y|$ corresponds to the total number of topics evaluated.

Uniqueness: The function $\sigma(\cdot)$ (in Equation (17)) is instantiated using the function $uniqueness(t_i, t_j)$ to define DU_{inter} as:

$$DU_{inter} = \frac{\sum_{t_i \in \mathbb{T}_z} \sum_{t_j \in \mathbb{T}_y} uniqueness(t_i, t_j)}{|\mathbb{T}_z| * |\mathbb{T}_y|} \tag{18}$$

SHS: The cosine distance among the centroids (Equations (12) and (13)) is exploited Equation (17). The semantics in terms of SHS inter-topic distance metrics are defined (DC_{inter}) as:

$$DC_{inter} = \frac{\sum_{t_i \in \mathbb{T}_z} \sum_{t_j \in \mathbb{T}_y} SHS(t_i, t_j)}{|\mathbb{T}_z| * |\mathbb{T}_y|} \tag{19}$$

Table 1
Dataset characteristics.

Dataset	#Feat	#Doc	Density
WhatsApp	1,777	2,956	3.103
Angrybirds	1,903	1,428	7.135
Evernote	6,307	8,273	11.002
Dropbox	2,430	1,909	9.501
Pinterest	2,174	3,168	4.478
Tweets	8,029	12,030	4.450
Uber	5,517	11,541	7.868
InfoVis-Vast[8]	6,104	909	86.215
TripAdvisor	3,152	2,816	8.532
Facebook	5,168	12,297	6.427
ACM	16,811	22,384	30.428
20NewsGroup[9]	29,842	15,411	76.408

5. Experimental Setup

In this section, we present the experimental evaluation of our work. Section 5.1 presents the datasets used to perform the experiments. Section 5.2 shows the experimental setup. Section 6.1 presents the efficacy of the topic evaluation, considering the two standard topic quality measures: NPMI and Coherence. Finally, in Section 6.2, we evaluate the efficacy of the topic evaluation considering the three topic quality metrics proposed in Section 4.

5.1 Datasets

We consider 12 real-world datasets as a reference to evaluate the topic model quality of the HTM methods. Ten were comments collected in the Google Play Store. The Facebook and Uber datasets were proposed by Viegas et al. (2018), while the others were proposed by Guzman and Maalej (2014). The tweets dataset was proposed by Li et al. (2016). The ACM (Cunha et al. 2021) and 20NewsGroup are datasets exploited in several NLP works in the literature. These datasets were widely used in topic modeling (Viegas et al. 2018, 2019) and hierarchical topic modeling work (Viegas et al. 2020). Table 1 provides a summary of the reference datasets, reporting the number of features (words), documents, the mean number of words per document (density), and the corresponding references.

5.2 Algorithms and Procedures

We compare three variants of the proposed c-CluHTM with three HTM baselines: f-CluHTM, HLDA, and HPAM. f-CluHTM[10] is the original CluWords-based method that uses the pre-trained fastText embedding[11] to build the semantic matrix, described

8 https://www.cc.gatech.edu/gvu/ii/jigsaw/datafiles.html.
9 http://qwone.com/~jason/20Newsgroups/.
10 https://github.com/feliperviegas/cluhtm.
11 https://dl.fbaipublicfiles.com/fasttext/vectors-english/wiki-news-300d-1M.vec.zip.

in Section 3. The three c-CluTHM variants are part of the contribution of the work and use BERT's contextual embeddings to build the semantic matrix, as described in Section 3.4. As previously mentioned, we have three transformations of BERT's embedding to evaluate in the c-CluHTM method: (i) c-CluHTM (Max emb.), using BERT Max representation; (ii) c-CluHTM (Avg. emb.), using BERT Average representation, and (iii) c-CluHTM (Concat. emb.), using BERT Concatenate representation.

We run the HLDA and HPAM methods using the Mallet package.[12] For the HLDA, we empirically set $alpha = 10.0$, $\gamma = 0.1$, and $\eta = 0.1$. For HPAM, we set the topic balance and topic smooth to 1.0. Regarding the hARTM, we set to 1,000 the number of documents passed, the parent model weight (to build the hierarchy structure) to 0.90, and the regularizers SparsePhi and DecorrelatorPhi to 0.25 and $2.5e + 5$, respectively. For BERTopic, we used the author's source code.[13] In terms of parametrization, we exploited the BERT model to get the CLS embeddings using the same number of topics mentioned before.

We considered the top 10 words selected by the hierarchical topic modeling methods to represent the topics in all experimental evaluations of Section 6.1. In the experimental evaluation in Section 6.2, we considered 5 and 10 words to represent the latent topics. We also considered two levels of topologies to perform the topic evaluation. We assess the statistical significance of our results by exploiting a t-test with Bonferroni correction 95% confidence, the same statistical test exploited in Viegas et al. (2019, 2020).

6. Results

6.1 Experimental Evaluation of the Standard Topic Quality Metrics

This section evaluates the two main topic quality metrics used in the literature: NPMI and Coherence. As mentioned in Section 2.2, these metrics assess different topical quality aspects complementary to this assessment. The goal of this section is to evaluate whether the results in terms of topic quality are convergent among the evaluated metrics and then whether they are complementary concerning the results obtained with the new proposed metrics (Section 6.2).

Considering the NPMI results, we can observe in Figure 4 that the c-CluHTM variants achieved the best results in all evaluated datasets. The three c-CluHTM variants (Max, Avg, and Concat) tied with each other in all but two datasets (Pinterest and TripAdvisor), in which c-CluHTM (Max emb.) was a bit worse than the other two variants. c-CluHTM (Concat. emb.) achieves the best NPMI results in seven out of 12 datasets, tying with f-CluHTM in the following datasets: Evernote, Uber, InfoVis-Vast, ACM, and 20News. Regardless of static or contextual embeddings, the NPMI results show the superiority of all CluHTM solutions over HPAM, HLDA, BERTopic, and hARMT. Indeed, for some datasets, such as InfoVis-Vas, ACM, and 20News, the c-CluHTM results are remarkable, reaching NPMI scores close to 1.0 (the maximum).

Figure 5 shows the results for the Coherence metrics (described in Section 2.2). Similar behavior for the c-CluHTM variants can be observed as in the case of NPMI. The three c-CluHTM variants achieved the best results in all evaluated datasets. c-CluHTM

12 `https://github.com/mimno/Mallet`, a Java-based package for statistical natural language processing, document classification, clustering, topic modeling, information extraction, and other machine learning.

13 `https://maartengr.github.io/BERTopic/index.html`.

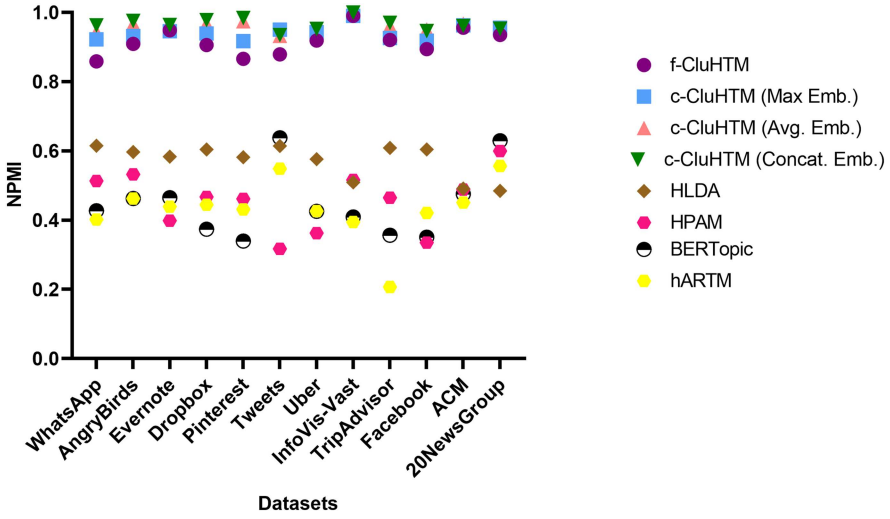

Figure 4
NPMI topic evaluation. c-CluHTM (Avg. emb.) and c-CluHTM (Concat. emb.) produce the best results in all evaluated datasets.

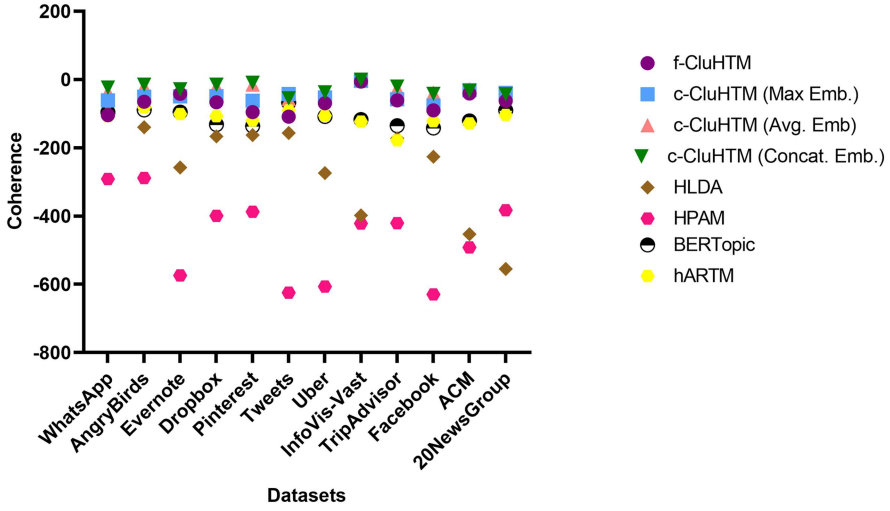

Figure 5
Coherence topic evaluation. c-CluHTM (Max emb.), c-CluHTM (Avg. emb.), and c-CluHTM (Concat. emb.) are the best solutions achieving the best results in all the datasets evaluated.

(Concat. emb.) produced the highest results, outperforming f-CluHTM in three out of 12 datasets, while the HLDA solution tied in the WhatsApp dataset.

Contrasting NPMI and Coherence metrics, we can see that the metrics captured different aspects of the topical quality, but both focus on measuring the quality of the topics based on the information shared among the words that compose the latent topics. Coherence has a higher sensitivity, generating variability between the topic scores. This is why, for example, for the AngryBirds dataset, c-CluHTM (Concat. emb.) is statistically equivalent to f-CluHTM. In addition, Coherence does not have an upper bound limit,

which makes each topic evaluation unique. Despite these differences, both NPMI and Coherence are consistent regarding the best methods to build topics.

Summarizing, the above results positively answer **RQ1** by showing that regardless of the transformation, using BERT contextual embeddings produced superior or equal results to those obtained with the static embeddings exploited in f-CluHTM in all datasets. The results also emphasize that even the use of static embeddings along with the f-CluHTM strategy can produce superior results to the other two evaluated representative HTM baselines.

The topic quality metrics considered in this section showed convergence regarding the best HTM methods, with some metrics, such as NPMI, almost achieving an upper limit in several datasets. The topic quality metrics presented in Section 4 were proposed to complement the evaluation of the topics. Therefore, they aim to evaluate other relevant aspects of the HTM scenario. We also want to verify whether the results of these newly proposed metrics (Uniqueness and SHS) can show new perspectives regarding the topic evaluation.

6.2 Experimental Evaluation of the Proposed Topic Quality Metrics

We consider for this evaluation c-CluHTM (Concat. emb.) (the best c-CluHTM variant in Section 6.1), f-CluHTM, HLDA, HPAM, and hARTM. BERTopic has a different way of building the topologies (as mentioned in Section 2.1). Since it did perform among the best methods in the previous evaluation, we did not consider it here. Regarding hARTM, it can only generate hierarchical structures for the ACM and 20NewsGroup, the largest datasets considered in our evaluation. This happens because hARTM requires considerable data to associate a probability distribution with latent topics over a hierarchical structure. We kept this method in our new evaluation, comparing it against the other methods only in ACM and 20NewsGroup.

We assess the statistical significance of the results with a t-test with Holm-Bonferroni correction with 95% confidence. This test assures that the best results, marked with a green triangle (▲), are statistically superior to all others. Statistical ties are represented as a yellow dot (●).

Table 2
Uniqueness intra-topic analysis considering five words per latent topic.

Datasets	c-CluHTM (Concat. emb)	f-CluHTM	HLDA	HPAM	hARTM
WhatsApp	**0.9996** ●	**0.9984** ●	0.6826	**0.9947** ●	–
AngryBirds	**0.9983** ●	**0.9995** ●	0.8896	**0.9613** ●	–
Evernote	**1.0000** ●	**1.0000** ●	**0.9501** ●	**0.9525** ●	–
Dropbox	**0.9989** ●	**0.9958** ●	**0.9309** ●	**0.9784** ●	–
Pinterest	**0.9989** ●	**0.9979** ●	**0.9456** ●	**0.9804** ●	–
Tweets	**1.0000** ●	**0.9997** ●	**0.9118** ●	**0.9982** ●	–
Uber	**1.0000** ●	**1.0000** ●	**0.9653** ●	**0.9247** ●	–
InfoVis-Vast	**0.9943** ●	**0.9996** ●	**1.0000** ●	**0.9800** ●	–
TripAdvisor	**0.9980** ●	**0.9999** ●	**0.9227** ●	**0.9676** ●	–
Facebook	**0.9995** ●	**0.9947** ●	0.9222	**0.9827** ●	–
ACM	**1.0000** ●	**1.0000** ●	**0.9945** ●	**0.9838** ●	**0.9875** ●
20NewsGroup	**0.9934** ●	**1.0000** ●	**0.9971** ●	**0.9316** ●	**0.9800** ●

Table 3
Uniqueness intra-topic analysis considering ten words per latent topic.

Datasets	c-CluHTM (Concat. emb)	f-CluHTM	HLDA	HPAM	hARTM
WhatsApp	**0.9973** ●	**0.9983** ●	0.6041	**0.9914** ●	–
AngryBirds	**0.9949** ●	**0.9989** ●	0.8115	**0.9722** ●	–
Evernote	**1.0000** ●	**1.0000** ●	0.8934	**0.9551** ●	–
Dropbox	**0.9979** ●	**0.9927** ●	0.8748	**0.9827** ●	–
Pinterest	**0.9964** ●	**0.9958** ●	0.8690	**0.9844** ●	–
Tweets	**1.0000** ●	**0.9995** ●	0.7871	**0.9979** ●	
Uber	**0.9998** ●	**1.0000** ●	0.8990	**0.9426** ●	–
InfoVis-Vast	**0.9926** ●	**0.9976** ●	**0.9964** ●	**0.9792** ●	–
TripAdvisor	**0.9985** ●	**0.9998** ●	0.8457	**0.9710** ●	–
Facebook	**0.9989** ●	**0.9961** ●	0.8243	**0.9790** ●	
ACM	**0.9988** ●	**1.0000** ●	**0.9771** ●	**0.9818** ●	**0.9865** ●
20NewsGroup	**0.9913** ●	**0.9975** ●	**0.9971** ●	**0.9361** ●	**0.9800** ●

6.2.1 Evaluation in Terms of Uniqueness. Tables 2 and 3 present the *intra-topic distance* in terms of the Uniqueness of topics for five and ten words per latent topics, respectively. Contrasting these two scenarios, we can observe that increasing the number of words representing the latent topics increases the sensitivity of the Uniqueness score. We can also observe that most methods achieved a statistical tie in most datasets, but HLDA performed poorly in some datasets. Focusing on the results of HLDA, we can observe that the number of words changed the quality of the method regarding the duplicity of words in the latent topics. This means this method starts to add duplicated words when it increases the number of words composing the latent topics. This interesting phenomenon can be captured by Uniqueness but not by traditional metrics.

Table 4 presents some topics generated by HLDA to show the behavior captured by the Uniqueness metric in the WhatsApp dataset, considering ten words. We can

Table 4
Example of topic topology built for the WhatsApp dataset using HLDA.

Main topic	Subtopics
ultimate hour tool verification received community region relative shift expanding	bug community region relative shift expanding bye news brought seamlessly
	code constant reliable community region relative shift expanding bye news
	community region relative shift expanding bye news brought seamlessly displays
	community region relative shift expanding bye news brought seamlessly displays
access loads talk androids community region relative shift expanding bye	community region relative shift expanding bye news brought seamlessly displays
	documents service community region relative shift expanding bye news brought
	community region relative shift expanding bye news brought seamlessly displays
	community region relative shift expanding bye news brought seamlessly displays
	community region relative shift expanding bye news brought seamlessly displays

Table 5
Example of topic topology built for the WhatsApp dataset using c-CluHTM (Concat. emb.).

Main topic	Subtopics
displayed requested opens releases blinking managed generated alternatives designed issues	blank activate prompt breach insert shift blinking center react releases
	glaring generated revelations shuffled democratic tempo expects arrived hooked listened
	format speak recorder sound picture texts bubble wall papers compression
	entering custom internal usable default locally lightweight ubiquitous utilizing implemented
usable windows default virus implemented pack operating offer development bugs	massive cache stored storage bean menu prompt feed screen rights
	beats hip download label viral annual exceed chill compression recorder
	compulsory unable unknown fulfilling possibly including preferred enhanced tempo entered
	react breach prompt loads implemented enhance factory custom reset releases
	volleyball functions medicine mice ken implemented providers carrier nickel japan

observe in the table that there are several duplicated subtopics (in **bold**) in the topology with basically the same words. This justifies the low Uniqueness score obtained for this method according to Equation (15). In contrast, Table 5 shows some latent topics generated by c-CluHTM (Concat. emb.). We can observe that there are no duplicated topics, justifying the high score of Uniqueness.

Traditional metrics assess the quality of the topics independently, and the final score is the average of the evaluated topics in isolation. These measures do not capture relationships among the subtopics based on their shared words, being unable to capture phenomena such as duplicated sub-topics illustrated in Table 4. In the example shown in Table 4, the subtopic *community region relative shift expanding bye news brought displays seamlessly* is repeated five times and, for the sake of calculation of NPMI and Coherence, this duplicated topic is also considered five times in the final score. This duplication will unavoidably cause distortions in the metric's final value.

Table 5 shows some topics built by c-CluHTM (Concat. emb.) in contrast to the topics built by HLDA. We can observe that the c-CluHTM (Concat. emb.) does not include duplications in the topics. It generated topics with some correlation regarding meaning/semantics (underlined words) and an outlier subtopic (marked in italics).

Tables 8 and 9 show the Uniqueness results in terms of *inter-topic distance*, regarding five and ten words per latent topic, respectively. We observe a similar behavior as in Tables 2 and 3—most of the methods are tied against each other, and the results for ten words show a decrease in the HLDA performance.

We performed a comprehensive manual analysis of the topics generated by each method, and HLDA built only duplicated topics. This showed that the Uniqueness method works, and most importantly, this method is capable of revealing issues regarding the generated topics. Indeed, Uniqueness can also be seen as a bias indicator for other evaluation metrics, such as NPMI and Coherence. For instance, in WhatsApp, the latent topic built by HLDA, shown in Table 4, *community region relative shift expanding*

Table 6
Example of topic topology built for the Facebook dataset using c-CluHTM (Concat. emb.). The underlined words show some examples of words that have semantic relationships.

Main Topic	Subtopics
cellular broadband micro fiber android proxy windows hardware wireless mobile	intended received reported actual noticed accompanied listed previously permanent written
	proxy worm port hosting authentication gateway windows administrators plug interfaces
	subway troll nexus toad sonata bike dong blackberry cord brother.
	mega bulk unit micro consumption hardware floppy tank cal chi
	pornographic stills recordings targeted photographs videos images artwork photography definition

bye news brought seamlessly displays has a score of around 0.58 for NPMI, and the average score of NPMI for HLDA in the WhatsApp dataset is around 0.61 (standard deviation of 0.03). This suggests that this latent topic is adding a bias to the final method's score.

6.2.2 Evaluation in Terms of SHS. We present *intra-* and *inter-topic distance* analyses, in terms of semantic information, as described in Section 4. As mentioned, this topic evaluation exploits *word embedding* vectors to capture the semantic relationship among words through cosine distance. Tables 10 and 11 present the results of *intra-topic distances* for five and ten words. Observing the results for five words, c-CluHTM (Concat. emb.) and f-CluHTM presented the best results, f-CluHTM achieved the best results in two datasets (Tweets and Facebook), meaning that the topics built by this method, considering the same topology, have higher semantic distances than the topics constructed by the other HTM methods. Regarding the results for ten words, f-CluHTM presented the best results in six out of 12 datasets. Interestingly, the number of words composing the latent topics increases the semantic information of the topics for f-CluHTM. Another relevant point is that HPAM presented the worst results for NPMI and Coherence, but for the InfoVis-Vast and 20NewsGroup datasets, HPAM statistically tied with c-CluHTM (Concat. emb.) and f-CluHTM.

Table 6 shows another example of the topic topology produced by c-CluHTM (Concat. emb.) for the Facebook dataset. We can see that the words *cellular* and *mobile* derive a topic that has the words *nexus* and *blackberry*, which are types of mobile phones. This topology also shows an interesting case, where the words *proxy* and *windows* are repeated in the main topic and second subtopic. However, the second subtopic is a more specific topic that "elaborates" why these words appeared in the main topic. Table 7 shows an example of the topic topology produced by f-CluHTM for the same Facebook dataset. Interestingly, the main topic refers to mobiles and bugs, and the second subtopics followed the same pattern by adding new words. For instance, the second and third topics are more related to bugs, while the others have more mobile-related words. We can also observe that words more semantically related to the main

Table 7
Example of topic topology built for the Facebook dataset using f-CluHTM. The underlined words show some examples of words that have semantic relationships.

Main Topic	Subtopics
handset cellphone helpline telephone mobiles blackberry cell bugged bugging battery	helpline timer tel mobiles clock handset blackberry bugging cellphone cell
	ELS cycle restarts timer rewind pace calls shutdown multiples retrying
	reinstalled reinstalls reinstalling autostart unpin reinstall installs tamper installing liquidate
	listening typing replying phoned talking chatting replies busy contacting speaking
	queries threads replies responses requests messages comments notices notifications remarks

Table 8
Uniqueness inter-topic analysis considering five words per latent topic.

Datasets	c-CluHTM (Concat. emb)	f-CluHTM	HLDA	HPAM	hARTM
WhatsApp	**0.9937** ●	**0.9969** ●	0.6889	**0.9696** ●	–
AngryBirds	**0.9927** ●	**0.9965** ●	0.9077	**0.9382** ●	–
Evernote	**0.9964** ●	**0.9950** ●	**0.9544** ●	**0.9327** ●	–
Dropbox	**0.9945** ●	**0.9976** ●	**0.9375** ●	**0.9546** ●	–
Pinterest	**0.9914** ●	**0.9976** ●	**0.9485** ●	**0.9568** ●	–
Tweets	**0.9935** ●	**0.9990** ●	**0.9111** ●	**0.9736** ●	–
Uber	**0.9975** ●	**0.9954** ●	**0.9656** ●	**0.9020** ●	–
InfoVis-Vast	**0.9892** ●	**0.9958** ●	**0.9943** ●	**0.9514** ●	–
TripAdvisor	**0.9936** ●	**0.9966** ●	**0.9285** ●	**0.9421** ●	–
Facebook	**0.9974** ●	**0.9976** ●	0.9298	**0.9560** ●	–
ACM	**0.9964** ●	**0.9952** ●	**0.9942** ●	**0.9591** ●	**0.9985** ●
20NewsGroup	**0.9963** ●	**0.9994** ●	**0.9977** ●	**0.9054** ●	**0.9745** ●

topics are at the bottom of the top words. For instance the first subtopic, the words *blackberry*, *bug*, *cellphone* are in the top 7, 8, 9, respectively. This explains why the results are better for ten words in Table 11.

Revising the topics of WhatsApp for HLDA (Table 4), we can also observe that the duplicated topics—*community region relative shift expanding bye news brought seamlessly displays*—impacted the results of SHS, as they have no semantic relationship with the *Main Topics* of Table 4.

Finally, Tables 12 and 13 show the results for *topic inter-distance*, considering five and ten words per latent topic, respectively. Again, f-CluHTM is the best method, tying with c-CluHTM (Concat. emb.) in two out of 12 datasets with five words. Similarly to what we infer from Tables 10 and 11, we can say that the topics built by f-CluHTM are more dissimilar in different topologies regarding semantic information. This reflects the desired behavior of a "good" HTM method since different topologies should express

Table 9
Uniqueness inter-topic analysis considering ten words per latent topic.

Datasets	c-CluHTM (Concat. emb)	f-CluHTM	HLDA	HPAM	hARTM
WhatsApp	**0.9888** ●	**0.9956** ●	0.6074	**0.9665** ●	–
AngryBirds	**0.9874** ●	**0.9945** ●	0.8235	**0.9474** ●	–
Evernote	**0.9942** ●	**0.9930** ●	0.8977	**0.9360** ●	–
Dropbox	**0.9914** ●	**0.9954** ●	0.8865	**0.9582** ●	–
Pinterest	**0.9871** ●	**0.9962** ●	0.8792	**0.9605** ●	–
Tweets	**0.9924** ●	**0.9985** ●	0.7863	**0.9729** ●	–
Uber	**0.9947** ●	**0.9945** ●	0.8995	**0.9194** ●	–
InfoVis-Vast	**0.9842** ●	**0.9913** ●	**0.9947** ●	**0.9510** ●	–
TripAdvisor	**0.9918** ●	**0.9950** ●	0.8506	**0.9448** ●	–
Facebook	**0.9958** ●	**0.9963** ●	0.8308	**0.9533** ●	–
ACM	**0.9952** ●	**0.9949** ●	**0.9776** ●	**0.9577** ●	**0.9963** ●
20NewsGroup	**0.9951** ●	**0.9990** ●	**0.9968** ●	**0.9098** ●	**0.9781** ●

Table 10
SHS intra-topic analysis considering five words per latent topic.

Datasets	c-CluHTM (Concat. emb)	f-CluHTM	HLDA	HPAM	hARTM
WhatsApp	**0.3657** ●	**0.3919** ●	0.2696	0.2879	–
AngryBirds	**0.3783** ●	**0.4082** ●	0.3017	0.3046	–
Evernote	**0.3666** ●	**0.4410** ●	0.3287	0.2997	–
Dropbox	**0.3932** ●	**0.3629** ●	**0.3587** ●	0.3084	–
Pinterest	**0.3535** ●	**0.4012** ●	0.3135	0.3001	–
Tweets	0.3886	**0.3950** ▲	0.3680	0.2987	–
Uber	**0.3856** ●	**0.4344** ●	0.3219	0.3082	–
InfoVis-Vast	**0.3355** ●	**0.3190** ●	**0.2933** ●	**0.2868** ●	–
TripAdvisor	**0.3805** ●	**0.4151** ●	0.3504	0.3072	–
Facebook	0.3756	**0.4532** ▲	0.3409	0.2994	–
ACM	**0.3793** ●	**0.3990** ●	**0.3767** ●	**0.3267** ●	0.2497
20NewsGroup	**0.3452** ●	**0.3932** ●	**0.3733** ●	**0.4183** ●	0.2728

distinct dataset concepts. f-CluHTM and c-CluHTM (Concat. emb.) were superior to the other baselines by considerable margins. It is important to note that since SHS measures the semantic distance among word vectors, the pooling transformation applied to transform the BERT embeddings into a static version may have impacted the cosine distance measure compared to the fastText embeddings since pooling strategies are susceptible to loose information.

However, when contrasting the SHS inter-topic with the NPMI and Coherence results, we can observe that the pooling does not affect the quality of the topics regarding the co-occurrence information among words. Again, the SHS results show another perspective on analyzing the quality of topics that goes beyond the traditional metrics.

6.2.3 Correlation Among Traditional and New Metrics. Finally, to further motivate the need for our new metrics and to confirm that they indeed capture different HTM properties, we performed a Pearson correlation analysis between the traditional HTM evaluation metrics and the newly proposed ones. For this analysis, we selected the results from the c-CluHTM (Concat. emb.) method. Figure 6 shows the heatmap of pairwise Pearson

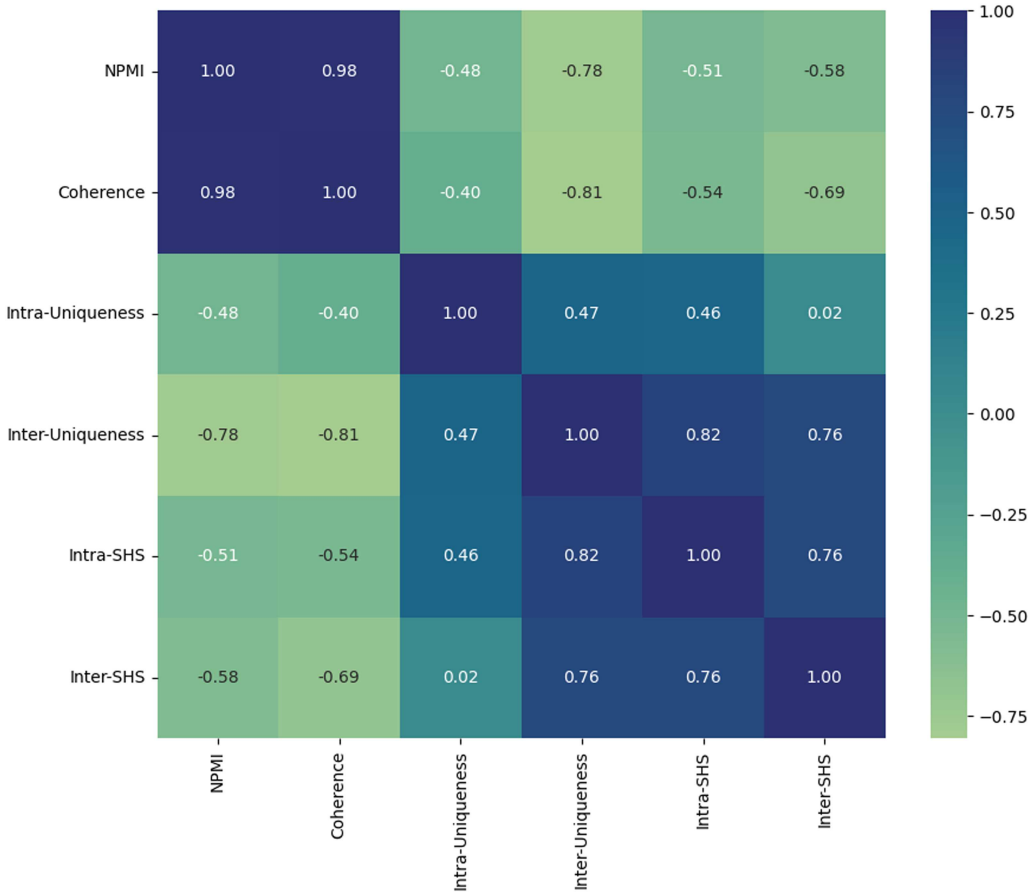

Figure 6
Pearson correlation between HTM evaluation metrics.

correlations. As can be seen, the proposed metrics usually have a low to moderate (negative) correlation with NMPI and Coherence. This result reinforces that the new metrics capture different HTM properties than the traditional ones, supporting their need.

6.3 Analysis of the Stability of the CluHTM Solution

In this section, we analyze the stability of the results produced by our solutions and proposed metrics. For this, we perform a new series of experiments with CluHTM using different initialization seeds to check whether this impacts the evaluation results. It is also important to note that all the CluHTM variants exploit the same stability measure (Greene, O'Callaghan, and Cunningham 2014), described in Section 3.2, as well as the NMF method to generate the latent topics, as described in Section 3. In a nutshell, the Stability measure executes S runs of NMF using different samplings of the dataset to select a fair number of topics for the dataset. In the original paper (Greene, O'Callaghan, and Cunningham 2014), the authors presented experiments considering eight datasets to show that the stability metric accurately chooses appropriate topics for the tested

Table 11
SHS intra-topic analysis considering ten words per latent topic.

Datasets	c-CluHTM (Concat. emb)	f-CluHTM	HLDA	HPAM	hARTM
WhatsApp	0.2538	**0.3207** ▲	0.1255	0.1955	–
AngryBirds	0.2674	**0.3379** ▲	0.1811	0.1912	–
Evernote	0.2760	**0.3892** ●	0.2012	0.2078	–
Dropbox	0.2845	**0.2961** ●	0.2099	0.2141	–
Pinterest	0.2462	**0.3257** ▲	0.1811	0.1872	–
Tweets	0.2699	**0.3319** ▲	0.2375	0.2033	–
Uber	0.2903	**0.3837** ▲	0.1975	0.2163	–
InfoVis-Vast	**0.2392** ●	**0.2597** ●	0.1707	**0.2063** ●	–
TripAdvisor	**0.2732** ●	**0.3557** ●	0.2196	0.2025	–
Facebook	0.2738	**0.3912** ▲	0.2109	0.2016	–
ACM	**0.3068** ●	**0.3471** ●	0.2574	0.2378	0.1708
20NewsGroup	**0.2770** ●	**0.3215** ●	0.2633	**0.2981** ●	0.2001

Table 12
SHS inter-topic analysis considering five words per latent topic.

Datasets	c-CluHTM (Concat. emb)	f-CluHTM	HLDA	HPAM	hARTM
WhatsApp	0.3755	**0.5070** ▲	0.2786	0.2883	–
AngryBirds	0.3993	**0.4983** ▲	0.3163	0.3030	-
Evernote	0.4148	**0.5041** ▲	0.3328	0.3078	–
Dropbox	0.4212	**0.4970** ▲	0.3787	0.3147	-
Pinterest	0.3819	**0.5027** ▲	0.3159	0.3138	–
Tweets	0.4354	**0.5471** ▲	0.3709	0.3005	–
Uber	0.4564	**0.5155** ▲	0.3224	0.3201	–
InfoVis-Vast	**0.3790** ●	**0.3671** ●	0.2950	0.2947	–
TripAdvisor	0.4013	**0.5009** ▲	0.3597	0.3071	–
Facebook	0.4295	**0.5510** ▲	0.3443	0.3138	–
ACM	**0.4899** ●	**0.5047** ●	0.3831	0.3422	0.2794
20NewsGroup	0.4938	**0.5909** ▲	0.3939	0.4308	0.3318

datasets against a consensus approach (Brunet et al. 2004) that creates connectivity of items appearing in the same topics after τ randomly NMF runs.

In our experimental evaluation, we select c-CluHTM (Concat. emb.) and complete three runs using different initialization seeds. We considered the statistical test described in Section 5 with ten words to compose the latent topics. Table 14 shows the results of the three runs, considering NPMI, Uniqueness (Intra- and Inter-), and SHS (Intra- and Inter-). We can observe that all three experiments statistically tie in all evaluation metrics, showing evidence that the seed parameter does not impact the performance of the HTM method. In addition, Table 15 illustrates some examples of latent topics retrieved from different runs. We can observe that the c-CluHTM (Concat. emb.) can generate similar topics for WhatsApp and ACM datasets and the same latent topics for the TripAdvisor dataset.

Table 13
SHS inter-topic analysis considering ten words per latent topic.

Datasets	c-CluHTM (Concat. emb)	f-CluHTM	HLDA	HPAM	hARTM
WhatsApp	0.2697	**0.4430** ▲	0.1289	0.1944	–
AngryBirds	0.2921	**0.4248** ▲	0.1881	0.1896	–
Evernote	0.3264	**0.4560** ▲	0.2046	0.2144	–
Dropbox	0.3119	**0.4273** ▲	0.2237	0.2187	–
Pinterest	0.2684	**0.4297** ▲	0.1852	0.1944	–
Tweets	0.3538	**0.4890** ▲	0.2387	0.2050	–
Uber	0.3638	**0.4661** ▲	0.1975	0.2261	–
InfoVis-Vast	0.2805	**0.3033** ▲	0.1721	0.2128	–
TripAdvisor	0.2920	**0.4413** ▲	0.2239	0.2060	–
Facebook	0.3321	**0.4915** ▲	0.2126	0.2110	–
ACM	0.4303	**0.4585** ▲	0.2620	0.2516	0.1914
20NewsGroup	0.4263	**0.5348** ▲	0.2818	0.3131	0.2501

Table 14
Analysis of the stability of c-CluHTM (Concat emb.) method.

Datasets	Runs	NPMI	Uniqueness Intra	Uniqueness Inter	SHS Intra	SHS Inter
WhatsApp	Run 1	0.9571 ●	0.9968 ●	0.9883 ●	0.2501 ●	0.2780 ●
	Run 2	0.9575 ●	0.9964 ●	0.9901 ●	0.2520 ●	0.2644 ●
	Run 3	0.9628 ●	0.9973 ●	0.9888 ●	0.2538 ●	0.2697 ●
TripAdvisor	Run 1	0.9677 ●	0.9987 ●	0.9918 ●	0.2640 ●	0.2999 ●
	Run 2	0.9672 ●	0.9961 ●	0.9928 ●	0.2633 ●	0.3000 ●
	Run 3	0.9700 ●	0.9985 ●	0.9918 ●	0.2732 ●	0.2920 ●
ACM	Run 1	0.9547 ●	0.9995 ●	0.9946 ●	0.3070 ●	0.4289 ●
	Run 2	0.9563 ●	0.9995 ●	0.9951 ●	0.2931 ●	0.4190 ●
	Run 3	0.9612 ●	0.9913 ●	0.9952 ●	0.3068 ●	0.4263 ●

6.4 Qualitative Analysis: Human Evaluation Experiment

We describe in this section the human evaluation experiment performed to further demonstrate, in a qualitative manner, the evaluative benefits of the new metrics. For this, we contrast two HTM approaches evaluated in the previous Sections, c-CluHTM (Concat. emb.) and HLDA. As observed, c-CluHTM (Concat. emb.) outperformed HLDA regarding both traditional HTM and the new proposed metrics (Uniqueness and SHS).

The experiments reported here aim to verify whether we can qualitatively produce similar results based on a human perception of the quality of the topics produced by both strategies. Accordingly, we selected the best two topics, in terms of NPMI score, for each HTM method for this evaluation. We stress that we should not directly compare the results of the qualitative and quantitative experiments as they use different data samples and principles.

To perform such (human) evaluation, we designed a Google Form, providing the best two topics and their two best-following sub-topics considering the five datasets

Table 15
Example of topics built by different initialization seeds.

Dataset	Run A	Run B
WhatsApp	interaction relationships sexual speech emails sites concept facilities carrier advertising	interaction sites texts carrier relationships pages compression forum sexual blackberry
	windows operating usable virus devices blackberry platforms annual development default	usable default windows custom annual virus viral offer operating internal
TripAdvisor	paint wake hold talk sketch end stand to draw word press	paint wake hold talk sketch end stand to draw word press
	mall swimming nature bird tree king downtown motor plaza countryside	mall swimming nature bird tree king downtown motor plaza countryside
ACM	programming compiler assembly script code oriented calculus function application programmer	function calculus semantics declaration abstraction programming oriented expressions statement formula
	programming oriented function compiler object logic code operating algebra computational	programming operating compiler software pascal java dos python assembly application

WhatsApp, Uber, Pinterest, Dropbox, and TripAdvisor. The participants had to compare the topics generated by the two HTM approaches to answer the following questions:

1. Considering that coherence is determined by the semantic correlation of the topic words, what is the level of coherence of the topics considering the highest level (parent) with regard to the sub-topics at the second level (children)?
2. Considering that redundancy is given by the repetition of topic words, what is the level of redundancy of the topics considering the highest level (parent) with regard to the sub-topics at the second level (children)?
3. Considering that coherence is determined by the semantic correlation topic words, what is the level of coherence among the sub-topics at the second level (children)?
4. Considering that redundancy is given by the repetition of topic words, what is the level of redundancy among the sub-topics at the second level (children)?

Each question had five rating levels (1 to 5), with 1 being less coherent (Questions 1 and 3) or very redundant (Questions 2 and 4) and 5 being very coherent (Questions 1 and 3) or less redundant (Questions 2 and 4). Both HTM strategies were randomly displayed as Strategy A and Strategy B for each dataset to avoid bias in the answer. All the topics selected to perform the test, as well as the individual results for each dataset, can be seen in Appendix Table A.1.

Seventeen participants, all Computer Science students (undergraduate, master's, and Ph.D. students), answered the four questions for the five datasets. Figures 7–10 show the overall results regarding the four form questions for the five evaluated datasets.

Figure 7 presents the overall results for Question 1. Over 81.7% of the participants answered that the main topic and their respective sub-topics built by the CluHTM are very cohesive. In comparison, less than 9.2% answered that the topics built by HLDA are very coherent. This evaluation aligns with the results of SHS, which strictly measures this behavior.

Figure 8 summarizes the responses for Question 2. Regarding the redundancy of the main topics and their respective sub-topics, most participants answered that both strategies built a hierarchy of topics with a low level of redundancy of words. Again, this result also aligns with our automatic experimental evaluation since we did not find redundancy at different hierarchy levels.

Responses for Question 3 are summarized in Figure 9, which contrasts the cohesion of the sub-topics. Over 56% of the participants answered that CluHTM has cohesive sub-topics, while 36.2% responded the same for the sub-topics built by HLDA. These results again align with the SHS results of previous quantitative experiments.

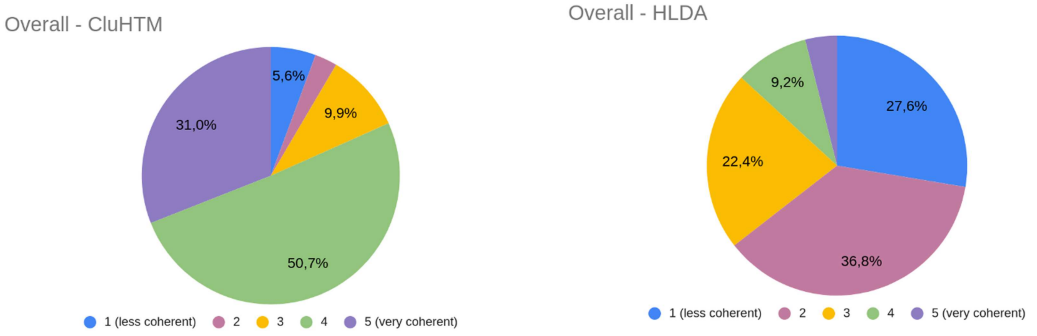

Figure 7
Summary results for Question #1: Answers vary from 1 to 5, with 1 being less coherent and 5 very coherent.

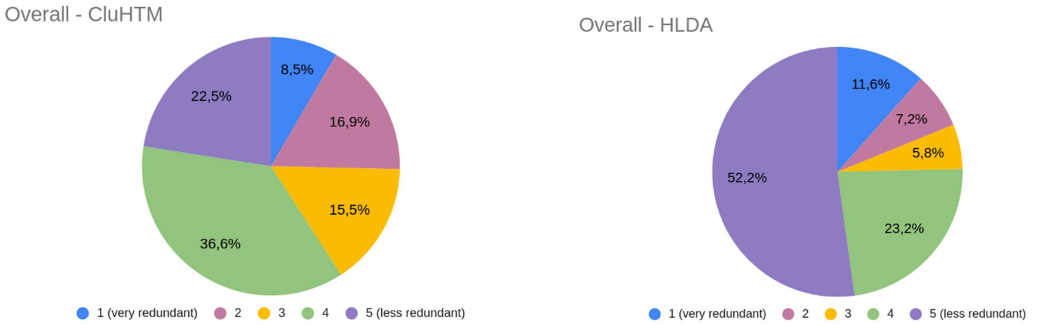

Figure 8
Summary results for Question #2: Answers vary from 1 to 5, with 1 being very redundant and 5 less redundant.

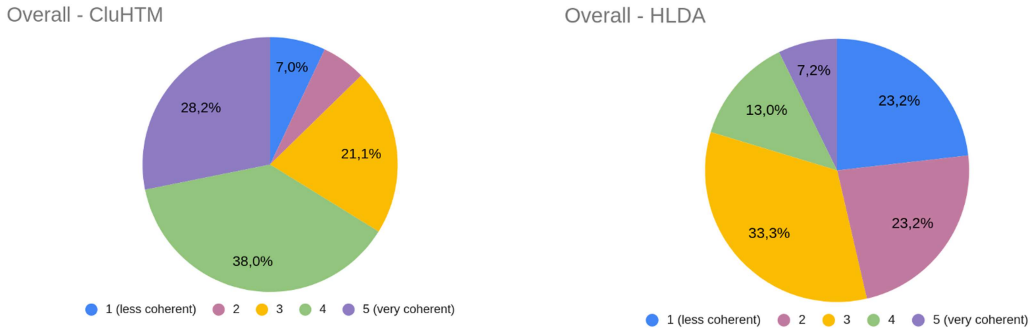

Figure 9
Summary results for Question #3. Answers vary from 1 to 5, with 1 being less coherent and 5 very coherent coherent.

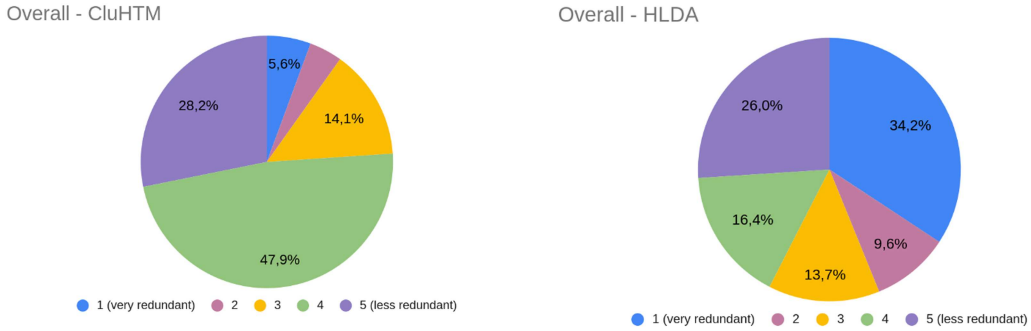

Figure 10
Summary results for Question #4. Answers can vary from 1 to 5, with 1 being very redundant and 5 less redundant.

Finally, Figure 10 shows the overall results for Question 4. The participants answered that CluHTM has less redundancy in the topics built in the second level (about 75%) when compared to HLDA (about 42%). In addition, 34.2% of the participants answered that topics built by HLDA are very redundant. These results are consistent with the Uniqueness results and the topic examples shown in the previous section. Particularly regarding Question 4, it is important to note that the duplicity of the subtopics does not occur in all the datasets, for instance, in the topics selected for the Uber dataset (see Appendix Table A.1).

In summary, we believe that the human evaluation is very consistent with the experimental results analyzed in the previous sections, especially regarding the proposed metrics, thus proving further confirmation for our hypotheses.

7. Discussion: Theoretical and Practical Implications and Limitations of our Study

In this section, we discuss theoretical and practical perspectives on the obtained results. We start by revisiting the research questions:

RQ1: *Does c-CluHTM build more informative topics when compared to f-CluHTM by exploiting contextual word embeddings?*

RQ2: *Are the concepts of intra- and inter-topic distance helpful in measuring the quality of the topic hierarchies built by HTM methods?*

From a theoretical point of view, our work revealed important aspects in evaluating HTM methods that are not captured by traditional topic modeling metrics. Indeed, most HTM proposals in the literature evaluate their methods using the same evaluation methodology explored in traditional topic modeling and do not consider some specific idiosyncrasies of the hierarchical aspect of the task. Our work constitutes a new milestone in the HTM evaluation methodology, demonstrating the need to assess the hierarchical structure of topics, especially issues related to the topology and redundancy of the hierarchical topics, in addition to assessing the topical quality using only traditional topic modeling metrics (NPMI and Coherence).

Our work considerably advances the knowledge in the field by focusing on the hierarchical structure during the evaluation, complementing and improving the current HTM evaluation methodologies. Indeed, the proposed metrics Uniqueness and SHS and the concepts of intra- and inter-distances demonstrated to be useful in providing perspectives and insights different from what can be supported by the traditional evaluation metrics, positively answering the **RQ2**. Our work also provides a better understanding of the role of semantics (i.e., context) in the HTM task.

From a practical point of view, our work affects not only how new HTM proposals should be evaluated in the future but also establishes a new state-of-the-art for the field with which new studies should be compared. Indeed, the CluHTM variants (c-CluHTM and f-CluHTM) achieved the best results reported in the literature for the datasets we experimented with, regarding both traditional and the newly proposed metrics, positively answering **RQ1**.

In our experimental evaluation, we observed that: (i) CluWords can capture semantic information from distinct embedding models; (ii) the transformation of contextual embeddings into a static embedding format can capture word semantic information for the metrics NPMI and Coherence. In summary, this work complements the original f-CluHTM work (Viegas et al. 2020), showing that the CluHTM solution can be effective for the HTM task, given the proper adaptations.

Despite its several contributions, our work has some limitations. The first ones are related to issues of the embedding representations we explore. f-CluHTM is built using pre-trained static embedding whose CluWords representation may suffer from a lack of context and an inability to capture the nuanced meaning of each word. In contrast, c-CluHTM is a variant that exploits modern transformers-based solutions. Nowadays, transformer-based solutions are trained in a greater variety of domains than static models. In addition, transformer-based models can better handle the tokens' coverage than static embeddings due to tokenizers' solutions (Byte Pair Encoding, WordPiece, SentencePiece, etc.). There is currently a plethora of transformer-based models trained for different domains, broadening the scope of application. However, c-CluHTM currently relies on pooling strategies, which may suffer from: (i) loss of information (some important information may be lost during the transformation) and (ii) difficulty in handling noisy data (some word-embedding may carry noise information,

and pooling may intensify it). Therefore, these limitations justify the lack of improvement between f-CluHTM and c-CluHTM (complementing the answer to **RQ1**). Despite the current CluWords' limitations (mentioned in Section 3.1) of only receiving as input static embeddings, they were designed to deal with some level of semantic noise by means of the filters, so both CluHTM strategies presented here produced good results in terms of the standard NPMI and Coherence metrics as well as new metrics proposed in this work. In any case, in the future, we intend to propose a new CluWords solution that better handles contextual embeddings.

Regarding the design of the desirable evaluation metrics for uniqueness-based distances, these metrics should ensure that they can effectively capture how distinct the topics are from each other. The intuition behind the Uniqueness metric relies on how the HTM strategies generate distinct topics. It was expected that most of the strategies would have higher scores since they all could generate distinct topics. For the SHS-based distances, the desirable properties should include the ability to measure the context information between the words that compose the topics. The intuition behind this metric was to build a measurement that captures the semantic information about the words using word embeddings and measures the contextual distance using cosine distance. This is particularly important given that most of the datasets are collections of reviews related to specific applications, which share crucial information about the application itself, even across topics of different hierarchies. This shared information is a key factor in our observations. Therefore, our proposed metrics do meet these criteria. In all cases, we observed smaller intra-topic distances than inter-topic distances, following our intuition about what we want the metrics to capture. This confirms that our metrics are effective in distinguishing between topics within the same hierarchy and across different hierarchies, thereby fulfilling the desirable properties for both uniqueness-based and SHS-based distances.

Nevertheless, one limitation regards the restriction to compare only two consecutive levels of a hierarchy at each evaluation. Regarding Uniqueness results, in Tables 2–3 and 8–9, especially for the HLDA method in the 10-word version (Tables 8 and 9), these results clearly show that there is a problem regarding redundancy in the topics generated by this method, which does not affect the other methods, thus the high values for them. This evidence revealed by the metric, motivated us to look further into this issue and indeed we found considerable repetition in the topics, intra- and inter-level. This redundancy evidence was not revealed by the traditional metrics. In these cases, the Uniqueness values for HDLA can get as low as 0.6.

Regarding the SHS metric, we have observed that most scores were generally low (around 0.3). Despite this, Tables 10–13 show that one method is clearly superior to the others in almost all datasets. Furthermore, if we look at the dataset level (lines), we can see differences of magnitude among the models that can achieve more than 100% in some cases. Even when comparing the best method (in case, f-CluHTM) against the runner-up method (c-CluHTM), differences in the SHS method in the same dataset (line) can achieve up to 40%. Thus we claim that SHS is really able to differentiate methods, emphasizing the best results. In any case, we will work on improving the discriminating power of the metrics in the future, for instance, by means of smoothing and normalization. However, as currently defined, we firmly believe they already help to fill a gap in the literature.

Finally, the metrics were not designed to evaluate hierarchical structures such as those produced by BERTopic, where the hierarchical structure is not well defined in levels of depth, as mentioned in Section 2.1. In any case, our results emphasize that the future body of research in the HTM field should promote the use of our new

metrics to enrich the experimental evaluation by considering topological (hierarchical) and redundancy issues, as well as continuing in the path of exploiting semantics in the HTM task. Our work also opens room for new developments regarding new ways of evaluating HTM methods.

8. Conclusion

In this work, we advanced the state-of-the-art in HTM by proposing an extension of the f-CluHTM (Viegas et al. 2020) method. c-CluHTM is a variant that exploits BERT's hidden layers using several pooling strategies to build a single-word embedding representation for each word in CluHTM's meta-word construction. Our extension overcomes limitations posed by static embeddings that cannot capture the contextual use of words, an essential aspect in building "good" topics, as revealed by our experiments.

As our second contribution, we proposed new evaluation metrics for evaluating hierarchical topic modeling methods. The newly proposed topic quality metrics assess aspects related to topological consistency (or redundancy) and the hierarchical semantic structure that are important to hierarchical methods. These are different and complementary aspects than those captured by traditional TM metrics such as NPMI and Coherence, which measure the quality of each topic in an isolated manner, disregarding topological relationships among topics.

We evaluate our proposals comparing with the state-of-the-art—f-CluHTM (the original proposal that exploits the pre-trained fastText embedding)—and four representative HTM baselines (HLDA, HPAM, hARTM, and BERTopic) considering 12 distinct and widely used data collections. Since our proposed metrics' goal is to complement the traditional topic quality metrics, we also include the two most traditional quality topic metrics in our experimental evaluation.

Our experimental evaluation showed that c-CluHTM outperforms the five baselines, sometimes by large margins, when considering the traditional metrics: Coherence and NPMI. Furthermore, our new proposed topic quality metrics were able to capture distinct behaviors from the topics built, including duplicity of topics built by some HTM methods. Our results also showed that c-CluHTM and f-CluHTM present the best results in building a hierarchical structure while avoiding redundancy.

In summary, the newly proposed topic quality metrics assess aspects related to topological consistency (or redundancy) and the hierarchical semantic structure that are important to hierarchical methods. These are different and complementary aspects than those captured by traditional Topic Modeling metrics such as NPMI and Coherence, which measure the quality of each topic in an isolated manner, disregarding topological relationships among topics. We believe that our proposed metrics reinforce the need to design evaluation metrics that consider other aspects of the HTM problem, such as the topic structure, that go beyond just the "quality" of the topic words, as measured by the traditional metrics such as NPMI. Our beliefs are supported by our extensive quantitative and qualitative experimental results with 12 datasets and four baselines, which help to support the generalization of our arguments.

Our proposals and results not only present a new perspective on topic quality metrics but also pave the way for exciting future research. For instance, they allow us to raise the question: Can we evaluate topics by considering multiple crucial factors for the problem simultaneously, such as word occurrence (i.e., Coherence) and semantic structure of the topic, and incorporate all these factors into a single metric? This general research question can potentially revolutionize the domain of Topic Modeling, not just hierarchical topic modeling.

We also believe that the proposed metrics can be used beyond topic modeling, such as hierarchical text classification. In this case, it is important, first, to evaluate how these metrics can be generalized. In this sense, as future work, we intend to perform a quantitative study deeply focused on understanding the behavior of the metrics using datasets such as 20News (news documents), where documents are labeled using a "natural taxonomy" or classification system defined by field experts. Another potential avenue of investigation is to understand the effect of temporal evolution of data (Mourão et al. 2008; Salles et al. 2010) on CluHTM and the proposed metrics in order to better comprehend how topics (and the relationships among words) evolve over time in datasets that span long periods of time.

Regarding CluHTM variants, we intend to adapt the CluWords structure to enable contextual word embedding representation without applying pooling techniques. Thus, it would be possible to exploit the BERT embeddings without transforming them into static ones. We intend to deal with this weakness of our approach by exploiting recent transformers-like contextual representations such as SPLADE (Sparse Lexical and Dense) (Formal et al. 2021). SPLADE presents a novel approach that effectively combines sparse and dense representations (Cunha et al. 2020) for enhanced document ranking and retrieval. SPLADE is designed to address the challenges posed by multi-token words in transformers-like models, which need to exploit pooling approaches to embed word vectors into a single one (i.e., c-CluHTM drawback). This is achieved through a transformer-based model that selectively activates relevant dimensions, ensuring that both lexical and semantic nuances are captured. Although potentially powerful, as far as we know, SPLADE-based representations have not been exploited in the context of topic modeling. Accordingly, in future work, we intend to adapt the c-CluHTM solution to exploit a SPLADE-based approach in the Clustering Step, allowing us to exploit contextual embedding without transforming it into a static embedding.

A. Human Evalution

This section presents more details regarding the human evaluation (Section 6.4). First, we present the topics selected for the human evaluation experiment, regarding the five datasets: WhatsApp, Uber, Pinterest, Dropbox, and TripAdvisor. As we can see in Figures A.1—A.5, for each dataset, we selected the best topic and its respective children for each HTM approach: c-CluHTM (Concat. emb.) and HLDA. For each dataset, we asked the following questions to the participants:

Question 1. Considering that coherence is given by the semantic correlation of the words that compose the topics. What is the level of coherence of the topics at the highest level (parent) in relation to the sub-topics at the second level (children)?

Question 2. Considering that redundancy is given by the repetition of words that compose the topics. What is the level of redundancy of the topics at the highest level (parent) in relation to the sub-topics at the second level (children)?

Question 3. Considering that coherence is given by the semantic correlation of the words that compose the topics. What is the level of coherence between the subtopics of the second level (children)?

Question 4. Considering that redundancy is given by the repetition of words that compose the topics. What is the level of redundancy between the subtopics of the second level (children)?

Table A.1
Quantitative results of all answers extracted from the human evaluation.

Dataset	Question	Answer									
		CluHTM					HLDA				
		1 (less coherent)	2	3	4	5 (very coherent)	1 (less coherent)	2	3	4	5 (very coherent)
WhatsApp	1	0	1	1	12	1	2	6	4	2	1
	2	0	2	1	8	4	7	5	1	1	1
	3	0	2	3	7	3	4	1	1	6	3
	4	0	2	2	4	7	13	0	0	0	2
Uber	1	0	1	0	8	5	9	9	2	1	0
	2	0	1	1	7	5	0	0	1	5	8
	3	1	1	1	6	5	6	6	2	0	0
	4	0	1	1	7	5	0	0	0	4	9
Pinterest	1	1	0	0	5	8	4	4	4	1	0
	2	2	8	3	0	1	0	0	1	5	7
	3	1	0	3	6	4	2	2	7	1	0
	4	1	0	3	9	1	5	5	3	4	0
Dropbox	1	2	0	2	3	7	4	4	4	0	1
	2	2	1	1	5	5	0	0	0	2	11
	3	2	1	3	4	4	1	1	8	2	2
	4	2	0	1	7	4	6	2	5	0	1
TripAdvisor	1	1	0	4	8	1	2	5	3	3	1
	2	2	0	5	6	1	1	0	1	3	9
	3	1	0	5	4	4	3	6	5	0	0
	4	1	0	3	7	3	1	0	2	4	7

Table A.1 shows the quantitative results of all the answers extracted from the human evaluation. This table was used to generate the overall results shown in Figures 7–10.

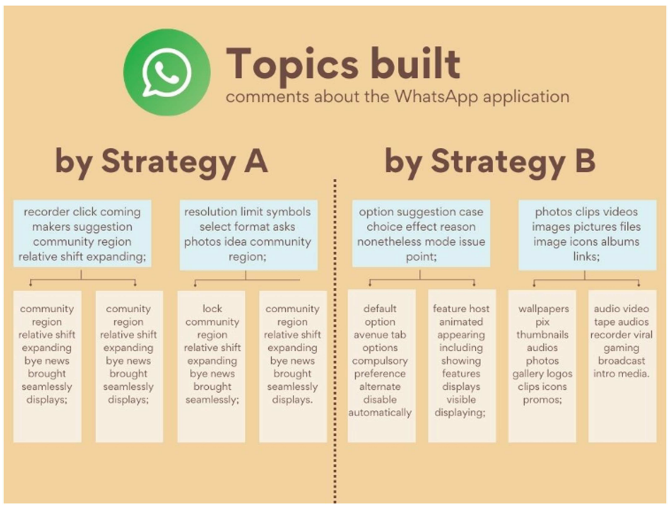

Figure A.1
Topics selected for the human evaluation, regarding the dataset WhatsApp.

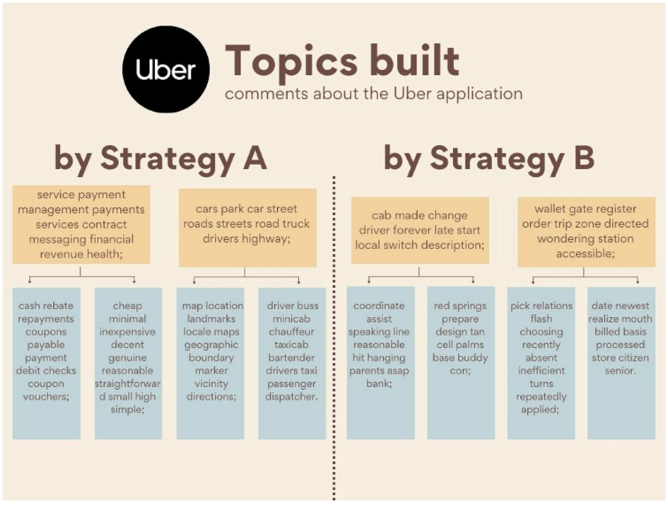

Figure A.2
Topics selected for the human evaluation, regarding the dataset Uber.

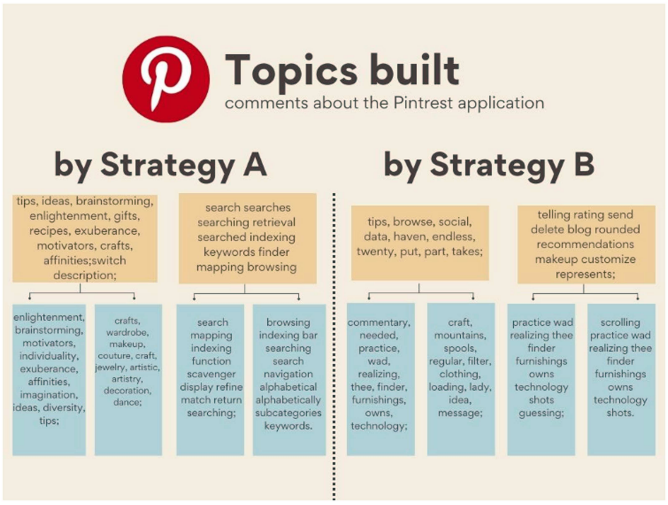

Figure A.3
Topics selected for the human evaluation, regarding the dataset Pinterest.

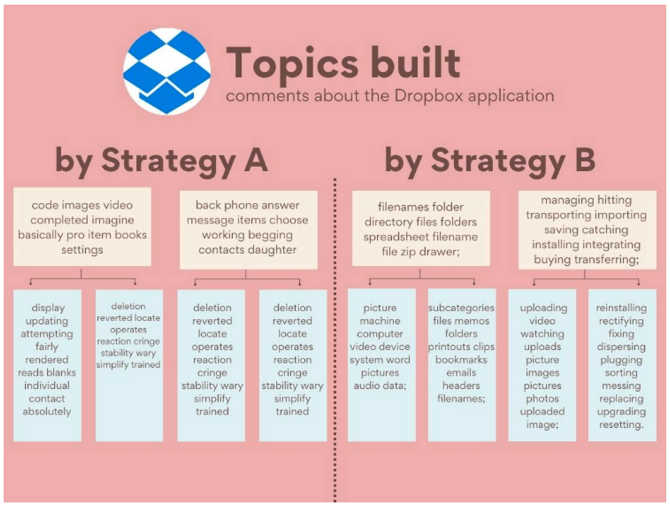

Figure A.4
Topics selected for the human evaluation, regarding the dataset Dropbox.

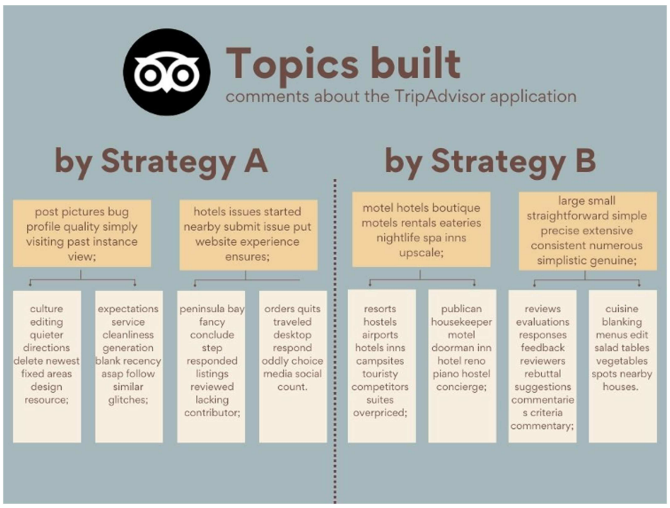

Figure A.5
Topics selected for the human evaluation, regarding the dataset TripAdvisor.

Acknowledgments

This work was supported by CNPq, CAPES, FAPEMIG, Amazon Web Services, NVIDIA, and Google Research Awards. All authors approved the final version of the manuscript.

References

Aziz, Saqib, Michael Dowling, Helmi Hammami, and Anke Piepenbrink. 2022. Machine learning in finance: A topic modeling approach. *European Financial Management*, 28(3):744–770. https://doi.org/10.1111/eufm.12326

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. 3:993–1022.

Bommasani, Rishi, Kelly Davis, and Claire Cardie. 2020. Interpreting pretrained contextualized representations via reductions to static embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781. https://doi.org/10.18653/v1/2020.acl-main.431

Brunet, Jean Philippe, Pablo Tamayo, Todd R. Golub, and Jill P. Mesirov. 2004. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*, 101(12):4164–4169. https://doi.org/10.1073/pnas.0308531101, PubMed: 15016911

Burkhardt, Sophie and Stefan Kramer. 2019. A survey of multi-label topic models. *SIGKDD Explorations Newsletter*, 21(2):61–79. https://doi.org/10.1145/3373464.3373474

Chauhan, Uttam and Apurva Shah. 2021. Topic modeling using latent Dirichlet allocation: A survey. *ACM Computing Surveys*, 54(7):1–35. https://doi.org/10.1145/3462478

Churchill, Rob and Lisa Singh. 2021. The evolution of topic modeling. *ACM Computing Surveys*, 54:1–35. https://doi.org/10.1145/3507900

Cunha, Washington, Sérgio D. Canuto, Felipe Viegas, Thiago Salles, Christian Gomes, Vítor Mangaravite, Elaine Resende, Thierson Rosa, Marcos André Gonçalves, and Leonardo Rocha. 2020. Extended pre-processing pipeline for text classification: On the role of meta-feature representations, sparsification and selective sampling. *Information Processing & Management*, 57(4):102263. https://doi.org/10.1016/J.IPM.2020.102263

Cunha, Washington, Vítor Mangaravite, Christian Gomes, Sérgio Canuto, Elaine Resende, Cecilia Nascimento, Felipe Viegas, Celso França, Wellington Santos Martins, Jussara M. Almeida, Thierson Rosa, Leonardo Rocha, and Marcos André Gonçalves. 2021. On the cost-effectiveness of neural and non-neural approaches and representations for text classification: A comprehensive comparative study. *Information Processing & Management*, 58(3):102481. https://doi.org/10.1016/j.ipm.2020.102481

de Andrade, Claudio M. V., Fabiano M. Belém, Washington Cunha, Celso França, Felipe Viegas, Leonardo Rocha, and Marcos André Gonçalves. 2023. On the class separability of contextual embeddings representations – or "The classifier does not matter when the (text) representation is so good!". *Information Processing & Management*, 60(4):103336. https://doi.org/10.1016/j.ipm.2023.103336

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Dieng, Adji B., Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453. https://doi.org/10.1162/tacl_a_00325

Formal, Thibault, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE v2: Sparse lexical and expansion model for information retrieval. arXiv:2109.10086. https://doi.org/10.1145/3404835.3463098

Greene, Derek, Derek O'Callaghan, and Pádraig Cunningham. 2014. How many topics? Stability analysis for topic models. In *Machine Learning and Knowledge Discovery in Databases*, pages 498–513, Springer. https://doi.org/10.1007/978-3-662-44848-9_32

Grootendorst, Maarten. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Guzman, Emitza and Walid Maalej. 2014. How do users like this feature? A fine grained sentiment analysis of app reviews. In *2014 IEEE 22nd International*

Requirements Engineering Conference (RE), pages 153–162. https://doi.org/10.1109/RE.2014.6912257

Han, Yue, Weihong Han, and Shudong Li. 2019. Review: Topic model application for social network public opinion analysis. In *Proceedings of the 2nd International Conference on Information Technologies and Electrical Engineering*. https://doi.org/10.1145/3386415.3386969

Li, Quanzhi, Sameena Shah, Xiaomo Liu, Armineh Nourbakhsh, and Rui Fang. 2016. TweetSift: Tweet topic classification based on entity knowledge base and topic enhanced word embedding. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management*, pages 2429–2432. https://doi.org/10.1145/2983323.2983325

Liu, Rui, Xingguang Wang, Deqing Wang, Yuan Zuo, He Zhang, and Xianzhu Zheng. 2018. Topic splitting: A hierarchical topic model based on non-negative matrix factorization. *Journal of Systems Science and Systems Engineering*, 27(4):479–496. https://doi.org/10.1007/s11518-018-5375-7

Liu, Yan, Ying Li, Chengcheng Hu, and Yongbin Wang. 2020. An method of improved HLDA-based multi-document automatic summarization of Chinese news. In *2019 6th International Conference on Dependable Systems and Their Applications (DSA)*, volume 1, pages 435–439. https://doi.org/10.1109/DSA.2019.00068

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. arXiv abs/1907.11692.

Malkov, Yu A. and Dmitry A. Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):824–836. https://doi.org/10.1109/TPAMI.2018.2889473, PubMed: 30602420

Mcauliffe, Jon and David Blei. 2007. Supervised topic models. In *Advances in Neural Information Processing Systems*, volume 20, pages 121–128.

Meng, Zaiqiao, Hong Shen, Huimin Huang, Wei Liu, Jing Wang, and Arun Kumar Sangaiah. 2018. Search result diversification on attributed networks via nonnegative matrix factorization. *Information Processing & Management*, 54(6):1277–1291. https://doi.org/10.1016/j.ipm.2018.05.005

Mikolov, Tomas, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Miles, Samuel, Lixia Yao, Weilin Meng, Christopher M. Black, and Zina Ben Miled. 2022. Comparing PSO-based clustering over contextual vector embeddings to modern topic modeling. *Information Processing & Management*, 59(3):102921. https://doi.org/10.1016/j.ipm.2022.102921

Mourão, Fernando, Leonardo Rocha, Renata Braga Araújo, Thierson Couto, Marcos André Gonçalves, and Wagner Meira Jr. 2008. Understanding temporal aspects in document classification. In *Proceedings of the International Conference on Web Search and Web Data Mining, WSDM 2008*, pages 159–170. https://doi.org/10.1145/1341531.1341554

Nikolenko, Sergey I. 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*, pages 1029–1032. https://doi.org/10.1145/2911451.2914720

Nikolenko, Sergey I., Sergei Koltcov, and Olessia Koltsova. 2017. Topic modelling for qualitative studies. *Journal of Information Science*, 43(1):88–102. https://doi.org/10.1177/0165551515617393

Perotte, Adler, Nicholas Bartlett, Noémie Elhadad, and Frank Wood. 2011. Hierarchically supervised latent Dirichlet allocation. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, pages 2609–2617.

Porturas, Thomas and R. Andrew Taylor. 2021. Forty years of emergency medicine research: Uncovering research themes and trends through topic modeling. *The American Journal of Emergency Medicine*, 45:213–220. https://doi.org/10.1016/j.ajem.2020.08.036, PubMed: 33059985

Resnik, Philip, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. 2015. Beyond LDA: Exploring supervised topic modeling for depression-related language

in Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 99–107. https://doi.org/10.3115/v1/W15-1212

Salles, Thiago, Leonardo Rocha, Gisele L. Pappa, Fernando Mour ao, Wagner Meira, Jr., and Marcos Gonçalves. 2010. Temporally-aware algorithms for document classification. In *SIGIR '10: Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 307–314. https://doi.org/10.1145/1835449.1835502

Shrivastava, Priya and Dilip Kumar Sharma. 2021. Fake content identification using pre-trained glove-embedding. In *2021 5th International Conference on Information Systems and Computer Networks (ISCON)*, volume 1, pages 1–6. https://doi.org/10.1109/ISCON52037.2021.9702379

Vayansky, Ike and Sathish A. P. Kumar. 2020. A review of topic modeling methods. *Information Systems*, 94:101582. https://doi.org/10.1016/j.is.2020.101582

Viegas, Felipe, Sérgio Canuto, Christian Gomes, Washington Luiz, Thierson Rosa, Sabir Ribas, Leonardo Rocha, and Marcos André Gonçalves. 2019. CluWords: Exploiting semantic word clustering representation for enhanced topic modeling. In *Proceedings of WSDM '19*, pages 753–761. https://doi.org/10.1145/3289600.3291032

Viegas, Felipe, Washington Cunha, Christian Gomes, Antônio Pereira, Leonardo Rocha, and Marcos Goncalves. 2020. CluHTM - semantic hierarchical topic modeling based on CluWords. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8138–8150. https://doi.org/10.18653/v1/2020.acl-main.724

Viegas, Felipe, Washington Luiz, Christian Gomes, Amir Khatibi, Sérgio Canuto, Fernando Mourão, Thiago Salles, Leonardo Rocha, and Marcos André Gonçalves. 2018. Semantically-enhanced topic modeling. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 893–902. https://doi.org/10.1145/3269206.3271797

Vorontsov, Konstantin, Oleksandr Frei, Murat Apishev, Peter Romov, and Marina Dudarenko. 2015. BigARTM: Open source library for regularized multimodal topic modeling of large collections. In *Analysis of Images, Social Networks and Texts*, pages 370–381. https://doi.org/10.1007/978-3-319-26123-2_36

Xu, Yueshen, Jianwei Yin, Jianbin Huang, and Yuyu Yin. 2018. Hierarchical topic modeling with automatic knowledge mining. *Expert Systems with Applications*, 103:106–117. https://doi.org/10.1016/j.eswa.2018.03.008

Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. Curran Associates Inc. https://doi.org/10.5555/3454287.3454804

Yu, Guanglei, Linlin Zhang, Ying Zhang, Jiaqi Zhou, Tao Zhang, and Xuehua Bi. 2022. Prediction and risk stratification from hospital discharge records based on hierarchical sLDA. *BMC Medical Informatics and Decision Making*, 22(1):14. https://doi.org/10.1186/s12911-022-01747-3, PubMed: 35033059

Zhou, Wei and Jelke Bloem. 2021. Comparing contextual and static word embeddings with small data. In *Conference on Natural Language Processing*, pages 253–259.