Revista de Informática Teórica e Aplicada - RITA - ISSN 2175-2745
Vol. 33, Num. 1 (2026) 87-98

RESEARCH ARTICLE

Exploring Topic Modeling in Short Texts from Social Media: a Comparative Analysis of Algorithms

Explorando a Modelagem de Tópicos em Textos Curtos de Mídias Sociais: uma Análise Comparativa de Algoritmos

Ian Macedo Maiwald Santos[1]*, Luciana de Oliveira Rech[1], Ricardo Moraes[2]

Abstract: Many texts are broadly disseminated on online social media platforms each day. Topic modeling is a Natural Language Processing and Unsupervised Learning technique used in this scenario. It identifies the topics in a collection of texts — that is, the most relevant groups of words in the context of all the texts are analyzed. The characteristics of the texts are a relevant factor for identifying topics. Unlike traditional sources, texts published on social media are usually short, because even when a character limit per publication (e.g., Twitter/X) is not imposed, users tend to be objective in the texts they write. This work evaluates the performance of four categories of topic modeling algorithms: traditional, Dirichlet Multinomial Mixture (DMM)-based, self-aggregation-based, and global co-occurrence-based. Real texts generated by social media users were used for the evaluation. Model performance was evaluated using quality metrics accepted in the literature. Finally, the results were analyzed such that the performance of each algorithm was weighted down, and to clarify whether there would be any detriment to the results of topic modeling using traditional algorithms on short texts.
Keywords: Topic Modeling — Short Texts — Social Media — Performance Comparison

Resumo: Diariamente, muitos textos são amplamente publicados nas redes sociais. A modelagem de tópicos é uma técnica de processamento de linguagem natural e de aprendizagem não supervisionada utilizada neste contexto. Esta técnica analisa os grupos de palavras mais relevantes em toda uma coleção de textos para identificar os tópicos. As características dos textos são um fator importante para a identificação de tópicos. Ao contrário de outras vias de publicação, os textos publicados nas redes sociais são normalmente curtos, porque mesmo quando não há limite de caracteres por publicação (por exemplo, no Twitter/X), as pessoas tendem a ser objetivas nos textos que escrevem. Este trabalho avalia o desempenho de quatro categorias de algoritmos de modelação de tópicos: Tradicional, baseado em Dirichlet Multinomial Mixture (DMM), baseado em auto-agregação e baseado em coocorrência global. Para a avaliação, foram utilizados textos reais gerados por usuários de redes sociais. O desempenho do modelo foi avaliado utilizando métricas de qualidade reconhecidas na literatura. Por fim, os resultados foram analisados para avaliar o desempenho de cada algoritmo, e esclarecer como o uso de algoritmos tradicionais em textos curtos afetam na modelagem dos tópicos.
Palavras-Chave: Modelagem de Tópicos — Textos Curtos — Mídias Sociais — Comparação de Performance

[1] *Departamento de Computação, Universidade Federal de Santa Catarina (UFSC), Brasil*
[2] *Departamento de Ciência da Informação, Universidade Federal de Santa Catarina (UFSC), Brasil*
***Corresponding author**: ian.mwld@gmail.com
DOI: http://dx.doi.org/10.22456/2175-2745.145569 • **Received:** 02/02/2025 • **Accepted:** 26/12/2025

1. Introduction

Social media have become increasingly popular in recent years and allow people to participate in online discussions and contribute opinions and information on various topics. With this significant participation, a large flow of data is generated by users and, as a result, online platforms are flooded with messages, mostly in the form of short texts [1]. However, some problems are more obvious in online discussions with many participants, while other trends are more discreet and go unnoticed [2]. For example, the COVID-19 pandemic highlighted the need to use large data sets for scientific purposes, as well as how this can help save lives [3].

In this context, topic modeling has emerged as a text mining technique that identifies semantic structures in a corpus of documents. For example, the Latent Dirichlet Allocation (LDA) developed by [4] is an unsupervised solution widely used to perform such modeling and can be applied in differ-

ent domains, such as political science [5], medical science [6], and identification of disinformation networks [2]. LDA is a statistical model that reveals the most prominent sets of semantically related words in a collection of texts, although it performs poorly when handling short texts [1]. Topic modeling demonstrates broad applicability across diverse domains, including the analysis of discourse, the examination of trends, or the investigation of public opinion dynamics, particularly in the context of elections [7].

However, the effective application of topic modeling solutions for social media faces several challenges. Social media texts are short in length, which makes it difficult to obtain efficient results when applying traditional topic modeling approaches such as LDA. This is mainly due to the sparsity of the data, as short texts provide limited context [8].

The effective application of topic modeling to social media texts is a challenge that has received increasing attention from the scientific community. To overcome the limitations of the traditional LDA approach in applications involving short documents, alternative models have emerged, such as Biterm Topic Modeling (BTM), which maps global patterns of word co-occurrence in the corpus. In addition, other categories of models have been proposed [9]. In this context, it is important to note that the application of these techniques in Portuguese, especially for social media, has been little explored in the scientific literature.

This study therefore evaluated the performance of topic modeling techniques in identifying topics of interest in short texts from social media, using the Portuguese language as a base. To achieve this goal, the performance of 16 topic modeling algorithms was evaluated with a corpus of Portuguese tweets related to the COVID-19 pandemic. Each technique was evaluated in terms of perplexity, topic distance, divergence, and topic coherence. The techniques were compared on 10 different scenarios to highlight the advantages and limitations of each.

Finally, this study sought to fill a gap in the scientific literature by evaluating the effectiveness of different topic modeling techniques on Portuguese-language social media texts. The results obtained are expected to contribute to the development of more accurate and reliable solutions for the analysis of large data sets generated on social media.

2. Theoretical Foundation

This section presents the concepts involved in this work as well as their theoretical basis.

2.1 Topic Modeling

Topic modeling is a statistical technique extensively employed in textual analysis to identify latent variables within large datasets, enabling the probabilistic representation of documents through word co-occurrence patterns [10]. The generated topics are relevant and concise formed by a set of semantically related words, making it easier to understand the concepts discussed in the documents. It plays a key role in current research related to natural language processing and makes it possible to analyze data from different sources and forms, often hierarchically classified [11].

This process uses statistical algorithms, such as LDA, to assign probabilities to the words in each document and to the words in each topic [10]. Initially, the algorithm randomly assigns words to topics and then iterates to adjust probabilities until it converges on a coherent topic distribution. The result is a representation of documents in terms of topics, where each topic is a set of semantically related words. This allows for more accurate and automated analysis of the main topics discussed in documents, which is helpful in understanding and organizing large sets of textual data.

A “word” or “term” represents the basic unit of individual data; a collection of words forms a “sentence”; a “document” represents one or more sentences consisting of N words; and a “corpus” represents a set of M documents, usually encompassing the entire data set. A “vocabulary” is the collection of all distinct words in a corpus. A “topic” characterizes the probability distribution that covers a given vocabulary.

2.2 Topic Modeling Categories

[9] list three main categories of topic modeling for short texts: models based on Dirichlet Multinomial Mixture (DMM) [12, 13], models based on global co-occurrence [14, 15], and self-aggregation methods [16, 17]. There are also traditional algorithms that are widely used for modeling topics in long texts, such as scientific articles, books, or reports. However, when applied to short texts, such as tweets or text messages, these algorithms can have limited performance due to the lack of context and word co-occurrence.

Traditional models consist of algorithms such as:

- Latent Dirichlet Allocation (LDA)
- Latent Semantic Analysis (LSA)
- Probabilistic Latent Semantic Analysis (PLSA)
- Non-negative Matrix Factorization (NMF)

Models based on DMM include:

- Dirichlet Multinomial Mixture (DMM)
- Generalized Polya urn Poisson DMM (GPU-PDMM)
- Negative sampling and Quantization Topic Model (NQTM)
- Laplacian DMM (LapDMM)

Self-aggregation methods are composed of algorithms such as:

- Self-Aggregation based topic model (SATM)
- Conditional Random Field regularized Topic Model (CRFTM)
- Pseudo-document-based topic modeling (PTM)
- Meta-Info Guided Aggregation (MIGA) model

Models based on global co-occurrence include:

- Biterm topic modeling (BTM)

- Word Network Topic Model (WNTM)
- Multiterm Topic Model (MTM)
- Noise Biterm Topic Model with Word Embeddings (NBTMWE)

This work evaluated the performance of 16 topic modeling algorithms divided into four categories: traditional, DMM-based, self-aggregation-based, and global co-occurrence-based.

2.3 Assessment Metrics

Topic modeling approaches make it possible to study the content of topics in a corpus of documents. However, interpreting the generated topics requires interpretative analysis, and choosing appropriate metrics to evaluate the performance of topic modeling algorithms can be challenging [18]. It is also common to use diverse datasets that are not representative of reality, which makes it difficult to compare algorithms. For this study, a real-world data set was used to define the scope of the corpus, where the quality of the topics generated must be evaluated. Intuitive human evaluation can be problematic in large datasets with many topics, so it is necessary to use quality criteria. Accuracy is an important criterion for analyzing the design possibilities of topic modeling algorithms [19]. Computational complexity and computation time are also important criteria [20]. Although the modeling process can be time-consuming in certain cases, it is rarely infeasible for this reason.

2.3.1 Perplexity

Evaluating topic modeling techniques is a challenging problem due to the unsupervised nature of these algorithms. [21] presented the perplexity approach, which is a computationally efficient and independent metric that measures the ability of a topic model to generalize. Perplexity assesses the model's uncertainty in predicting words in an unknown dataset and indicates how well the model can generalize to new documents. A model with lower perplexity is considered to be of higher quality, because it indicates a greater ability to predict words accurately. However, while perplexity is a useful metric for assessing quality, it does not necessarily reflect the human interpretability of the generated topics [22].

2.3.2 Topic Distance

The distance between topics plays a crucial role in topic clustering, where finding the appropriate number of topics is a challenge. The method proposed by [23] uses a density-based approach to determine the distance between topics. This method is based on corpus statistics and uses cosine similarity to measure the cohesion between texts. The density-based clustering is then applied to adjust the number of topics based on the density of topics to find the best topic structure.

When mapping the process of generating new topics, this method considers the words that connect different topics. The distance is determined by both the size of the data set and the sensitivity of the correlations inherent in the document collection. The density of each topic is calculated, which makes it possible to identify the most unstable topics and iteratively adjust the number of topics until a stable model is reached. A low distance between topics indicates semantic similarity; this may indicate redundancy or difficulty in properly separating different topics. A high distance, meanwhile, indicates semantic distinction and separation between topics, thus capturing different aspects and themes in the data. The preference for a high or low distance depends on the goals and specific characteristics of the data set, where well-separated topics may facilitate the identification of distinct themes, while a lower distance may be useful for subtheme analysis.

2.3.3 Symmetric Kullback–Leibler Divergence

The symmetric Kullback–Leibler (KL) divergence measure proposed by [24] is a metric that compares the dissimilarity between probability distributions on topic modeling. This measure uses the KL divergence, which quantifies the difference between two probability distributions. Symmetric KL divergence is calculated as the average of the KL divergences in both directions, providing a balanced and intuitive measure of dissimilarity. A high symmetric KL divergence indicates greater dissimilarity between the probability distributions of the topics, thus representing distinct topics with no significant overlap of terms. A low symmetric KL divergence indicates less dissimilarity, thus suggesting topics with similar characteristics. This metric is efficient and can be applied to different data sets, which allows the estimation of the appropriate number of topics based on the best accuracy in unseen data.

2.3.4 C_{umass} Coherence

The C_{umass} coherence metric presented by [25] is an automated measure to evaluate the quality of topics generated in topic modeling. This metric is based on the co-occurrence statistics of words in the corpus and does not depend on external knowledge. C_{umass} coherence is calculated as the average of the logarithm of the ratios between the co-occurrence of pairs of topic words and the frequency of each individual word. Higher C_{umass} coherence scores indicate greater coherence between topic words, thus reflecting more interpretable and useful topics, while lower scores may indicate less interpretable topics. This metric does not rely on semisupervised data or external reference sources, which makes it a useful tool for automated evaluation of topic quality [26].

2.3.5 C_{w2v} Coherence

The work of [27] analyzes and compares topics generated by different topic modeling algorithms on different corpus. They use coherence measures, including one based on distributional semantics, to evaluate the quality of topics. The C_{w2v} coherence metric is used; this is based on the semantic similarity between the most representative words of each topic, calculated with Word2Vec vectors. Higher C_{w2v} coherence scores indicate more coherent and meaningful topics, which makes it easier to interpret and identify the underlying theme. However, certain contexts may require lower coherence, including cases when documents are intentionally composed of

unrelated words or contain mixtures of different topics. [22] stress the importance of carefully reading the generated topics, regardless of the topic modeling technique used.

3. Methodology

Social media platforms such as Twitter/X play an important role as a venue for debate and expression of opinions on high-profile events. Rapid exchange of ideas and viral dissemination of information made Twitter a valuable source of real data. In this context, topic modeling is a widely used technique for extracting hidden information and identifying sets of relevant words in the texts posted by users. However, topic modeling in short texts, such as those found in microblogging platforms, presents specific challenges, such as the sparsity of text data. This study therefore sought to analyze different topic modeling algorithms for short texts originating from social media to contribute to the development of more efficient solutions in this context.

Four steps were required to analyze the topics of interest in COVID-19 discussions on Twitter/X:

1. collecting data from the platform;
2. selecting samples;
3. applying treatments to the selected documents;
4. applying the topic modeling algorithm.

These steps were taken to infer the topics present in the online discussions and to allow for a more in-depth analysis of the collected data.

In this study, a real data set was used to compare the topic modeling algorithms. Data were collected from Twitter and consisted of publications in Portuguese related to COVID-19. Data were collected based on terms, hashtags, and keywords associated with discussions on the topic. We chose Twitter as a data source due to its large collection of short texts generated by users and the possibility of collecting specific texts through hashtags and other indexing resources. Only tweets in Portuguese were selected, excluding retweets, from March 1, 2020, to March 1, 2021. In total, 16 algorithms (divided into four categories) were tested.

The dataset used in this study was derived from the large-scale curated collection introduced by [28], which comprises more than 1.12 billion COVID-19 related tweets. After language filtering and preprocessing, approximately 4 million tweets remained for use in our analysis.

Text preprocessing is a fundamental step in text analytics. In this step, the text was cleaned and normalized to remove noise and produce quality results. In the specific case of social media, such as Twitter, it is common to find unnecessary elements for topic modeling, such as user mentions and URLs. To achieve a good result, the preprocessing pipeline included converting all letters to lower case, filtering out unwanted elements such as user mentions and URLs, removing punctuation, stop words, and special characters. It also included applying lemmatization to reduce words to their radicals. In addition, frequency-based filtering was applied: words occurring fewer than 10 times in the corpus were excluded to avoid sparsity, while words appearing in more than 60% of the documents were removed to minimize the dominance of overly frequent terms. To further reduce noise, we employed an extended domain-specific stop word list tailored to Portuguese COVID-19 discussions. This list included colloquial abbreviations (e.g., pfv, mto, oq), common functional words (e.g., porque, enquanto, assim), and frequent verbs (e.g., ser, ter, pode, vai). The combination of frequency thresholds with domain-specific stop words ensured that the remaining vocabulary captured more meaningful and discriminative patterns for topic modeling. Lastly, all algorithms tests used the same pre-processed texts to guarantee comparable conditions.

All algorithms were tested under ten different configurations, in which the number of topics (K) ranged from 10 to 100, in steps of 10. For models such as LDA and DMM, which are known to be sensitive to prior hyperparameters (e.g., alpha and beta/eta), we adopted the default values provided by their libraries. No fine-tuning was performed, as the algorithms evaluated differ considerably in their assumptions and parameterization, which makes direct optimization across all models less straightforward. By relying on standard implementations, we ensured comparable conditions across categories. Nevertheless, this choice may have influenced the performance of certain algorithms, particularly those more dependent on hyperparameter calibration, and should be regarded as a limitation of the study.

All scripts used to preprocess the data and run the experiments are available in a public GitHub repository [1].

3.1 STATISTICAL ANALYSIS

To assess whether the observed differences among algorithm categories were statistically significant, we considered the use of non-parametric tests suitable for repeated measures. For each evaluation metric (perplexity, topic distance, symmetric KL divergence, C_{umass} coherence, and C_{w2v} coherence), category-level scores were computed at each value of K (10–100). These paired observations across K allowed the application of the Friedman test, which evaluates differences among multiple groups under matched conditions. When the Friedman test indicated significant differences ($p < 0.05$), post-hoc pairwise comparisons were performed using the Nemenyi test to identify which categories differed.

This procedure was chosen because the metrics under study do not necessarily follow normal distributions, and the algorithms evaluated differ considerably in their assumptions and parameterization, which makes direct parametric comparison less straightforward. The non-parametric approach therefore provided a robust framework for comparing categories while preserving the repeated-measures design across values of K.

[1] https://github.com/IanMaiwald/comparative-topic-modeling-short-text

4. Results and Discussion

This section presents the overall performance of the algorithms in a broader context by identifying the differences between the algorithms in terms of performance and how the parameter changes affected them. This is followed by an analysis of notable relationships between topic distance and symmetric KL divergence in some cases. Finally, general considerations about the results obtained in the study are presented.

4.1 Perplexity

Perplexity is a metric used to evaluate the quality of probabilistic language models, such as topic modeling algorithms. This measure indicates how efficient the model is at predicting words in an unknown data set, where a lower perplexity value corresponds to a higher-quality model.

Figure 1 shows an increase in perplexity as the number of topics in the model increased. This surge in perplexity may be indicative of a less accurate model fit, as the addition of more topics may lead to greater complexity in the model fit. However, the increase in perplexity could vary between different algorithms and categories.

Regarding the stability of the algorithms, it was found that all of them showed a stable increase in perplexity as the number of topics increased. However, certain algorithms, such as NQTM (based on DMM) and PTM (based on self-aggregation), showed a slower increase in perplexity compared to other algorithms in their respective categories. These results suggest that these algorithms may be more robust to changes in the number of topics.

When comparing the categories of algorithms, those based on self-aggregation and global co-occurrence tended to have lower perplexity values compared to traditional and DMM-based algorithms. This suggests that these types of algorithms may be better suited for short texts.

4.2 Topic Distance

Topic distance is a quantitative measure that evaluates the similarity between two topics in a multidimensional space. This metric can be used to analyze and compare the quality of topic models by determining the average distance between learned topics. Topic distance aids in the interpretation of topic modeling results, thus providing a clearer understanding of the latent structure discovered in the data and helping to identify redundancy or overlap between learned topics. A smaller distance between topics indicates that they are semantically similar, while a larger distance indicates that the topics are distinct and well separated.

In general, Figure 2 indicates that self-aggregation and global co-occurrence algorithms were likely to produce more distinct topics, with wider variations in distance between different test scenarios. In contrast, traditional and DMM-based algorithms showed lower distance values. They also proved to be more stable, with less variation in results.

Traditional algorithms use more classical approaches that share many similar features when generating topics. These structural similarities of the algorithms were reflected in the results, where they performed similarly in several cases. However, despite their greater sensitivity to model and parameter adjustments, those based on self-aggregation and global co-occurrence showed better performance — that is, greater distance between the topics.

4.3 Symmetric KL-Divergence

The symmetric KL divergence metric is a measure of the dissimilarity between two probability distributions. It is based on KL divergence, a non-symmetric measure that quantifies the difference between two distributions by calculating the additional information cost needed to approximate one distribution by the other. This property is particularly useful when comparing and evaluating models and distributions in different contexts, such as topic modeling.

As seen in Figure 3 a higher symmetric KL divergence value indicates greater dissimilarity between the probability distributions of the topics, where the topics are more distinct and address different topics without much overlap of terms. On the other hand, a lower divergence indicates less dissimilarity between the distributions, with topics sharing similar terms and patterns.

In general, algorithms based on self-aggregation and BTM (co-occurrence) tended to produce more distinct topics with higher divergence values. Traditional algorithms tended to have lower divergence values, while algorithms based on DMM and global co-occurrence (except BTM) had intermediate values.

In terms of stability, the traditional and DMM-based algorithms showed individual differences from the smaller algorithms in their divergence values. Although these algorithms individually showed stability, the category of traditional algorithms showed a considerable range of divergence. The self-aggregation algorithms also showed stability, but some had slightly more pronounced variation. In contrast, algorithms based on global co-occurrence showed greater variation in divergence values.

4.4 C_{umass} Coherence

C_{umass} coherence is a quantitative measure used to evaluate the quality of topics generated by topic modeling algorithms. It is based on the idea that coherent topics should contain words that co-occur frequently in the corpus. C_{umass} calculates the average of the logarithm of the ratios between the co-occurrence of word pairs in the topic and the frequency of each individual word. This metric is particularly useful for comparing and selecting topic models by helping to identify those that generate more interpretable and meaningful topics. Higher C_{umass} values indicate greater coherence between the words in the topics, thus suggesting that the algorithm has successfully captured relevant semantic and structural relationships.

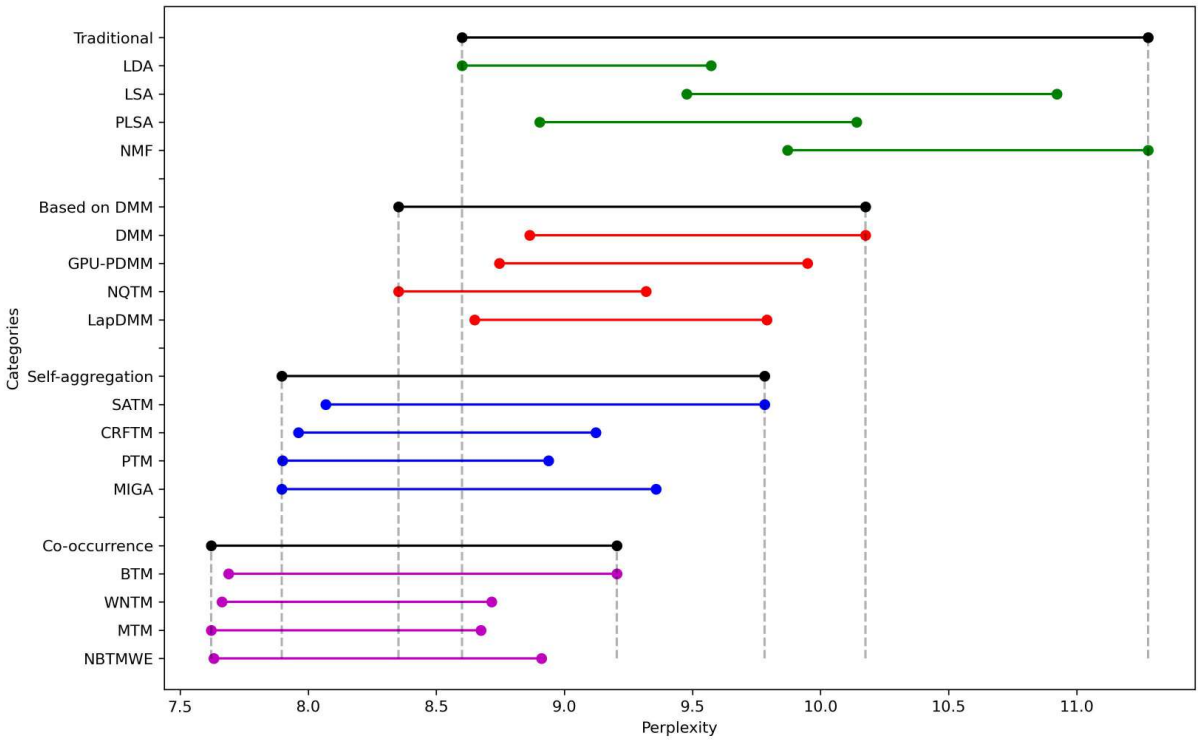

Figure 1. Result ranges for Perplexity. The colored lines represent the range of values obtained by varying K from 10 to 100.

It can be observed in Figure 4 that the algorithms based on global co-occurrence, especially BTM, showed significantly higher coherence values. The algorithms based on self-aggregation also showed good coherence values, while the traditional algorithms and those based on DMM showed lower values and were similar to each other.

In terms of stability, the traditional and DMM-based algorithms showed remarkable stability, with small variations in coherence values between different runs. On the other hand, the algorithms based on global co-occurrence showed a wide range of coherence values, thus indicating considerable instability and greater sensitivity to data and training conditions.

4.5 C_{w2v} Coherence

The C_{w2v} coherence metric is based on Word2Vec, a deep learning technique that generates vector representations of words based on the context in which they appear. In the context of topic modeling, C_{w2v} coherence quantifies the semantic similarity between the most representative words of a topic by calculating the average of the cosine similarities between the Word2Vec vectors of the words.

Topics with higher C_{w2v} values are considered more coherent and meaningful because their representative words have stronger semantic relationships, while topics with lower C_{w2v} coherence values are less coherent and may be more difficult to interpret.

In general, Figure 5 shows that topics generated by algorithms based on DMM and self-aggregation tended to have greater coherence and semantic relevance, with higher C_{w2v} values. Traditional algorithms and those based on co-occurrence generally had lower coherence values.

All algorithms showed some stability in their results, with small variations in the changes in the number of topics. However, certain algorithms, such as WNTM and NBTMWE, showed a slightly more pronounced variation in results, indicating instability compared to other algorithms.

These results highlight the effectiveness of the C_{w2v} metric in providing a unified view of the performance of algorithms in the same category that share common techniques for modeling topics.

4.6 FRIEDMAN AND NEMENYI TESTS

The Friedman test was applied across the four groups of algorithms (Traditional, DMM, self-aggregation, and Co-occurrence) for all five evaluation metrics, as shown in Table 1.

In each case, the test revealed statistically significant differences among the groups ($p* < 0.05$), indicating that the average performance was not equivalent across categories. For Perplexity and C_{umass} Coherence, the self-aggregation group

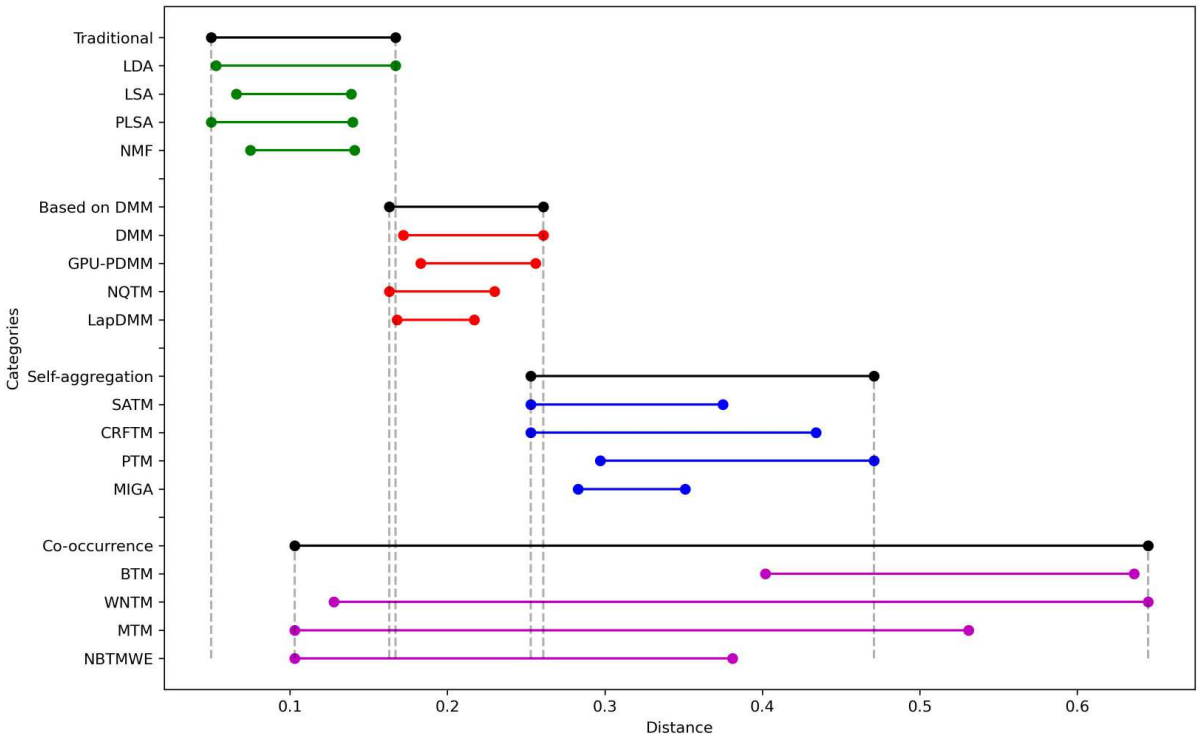

Figure 2. Result ranges for Distance. The colored lines represent the range of values obtained by varying K from 10 to 100.

Table 1. Summary of statistical significance tests (Friedman and Nemenyi) across evaluation metrics.

Metric	Friedman	p-value	Post-hoc summary
Perplexity	11.1	0.011	No strong pairwise differences
Topic Distance	10.8	0.013	No strong pairwise differences
Symmetric KL Divergence	11.8	0.008	Co-occurrence vs Traditional showed weaker similarity
C_umass Coherence	11.1	0.011	Auto-aggregation tended to diverge, but not strongly
C_w2v Coherence	11.0	0.012	No strong pairwise differences

showed a tendency to diverge more clearly from the others, with comparatively lower pairwise p-values when contrasted with Traditional and Co-occurrence methods. This suggests that models based on self-aggregation may produce outcomes that are somewhat more distinctive in terms of model fit and coherence, although the differences are not uniformly strong across all comparisons.

In the case of Symmetric KL-Divergence, the Co-occurrence group exhibited the most noticeable differentiation, particularly relative to Traditional algorithms, highlighting that co-occurrence-based models capture topic distributions in a manner that departs from more conventional approaches. For Topic Distance and C_{w2v} Coherence, although the Friedman test indicated overall significance, the post-hoc Nemenyi comparisons yielded p-values close to 1.0 across most pairs. This implies that, despite global differences, the groups perform in a relatively similar manner with respect to topic separation and semantic coherence.

Taken together, these findings suggest that while global differences exist across groups, the practical impact of algorithm selection is most pronounced in measures of model fit and distributional divergence. While acknowledging that pairwise differences are subtle and not always significant, the observed tendencies highlight the significance of matching algorithm selection with the particular evaluation criteria pertinent to the intended application.

4.7 Distance and Divergence

This study evaluated the performance of several topic modeling algorithms applied to short social media texts and used five metrics to measure the results.

Traditional algorithms showed low symmetric KL diver-

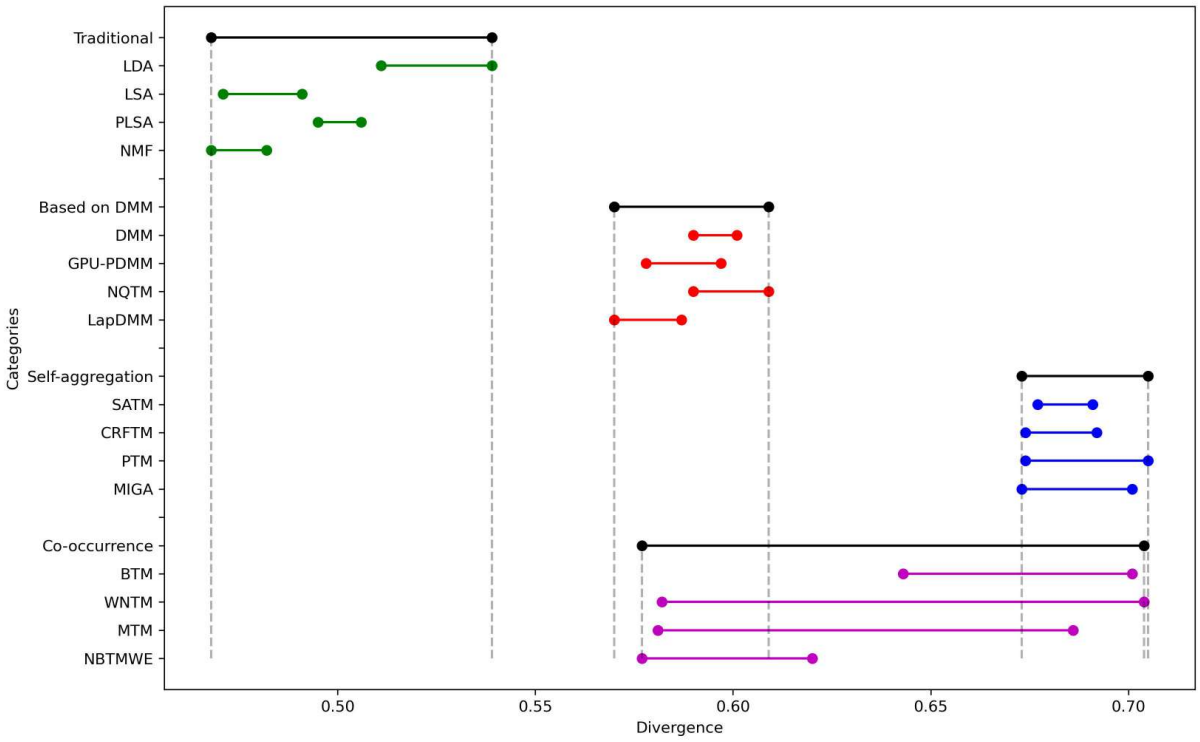

Figure 3. Result ranges for KL-Divergence. The colored lines represent the range of values obtained by varying K from 10 to 100.

gence values, thus revealing their limitations in generating topics with high probability distribution when applied to short texts. Topic distance was also not high, which indicated low to moderate topic differentiation. On the other hand, DMM-based algorithms showed higher divergences, indicating a greater ability to deal with the probability distributions of short texts.

Algorithms based on self-aggregation showed even larger divergences as well as high topic distances, which indicated a good balance between similarity and topic difference. Finally, algorithms based on global word co-occurrence showed mixed results for divergence and topic distance, although the BTM algorithm proved to be relatively stable and performed well in several tests.

In summary, it is important to consider multiple metrics when evaluating a topic modeling algorithm, rather than relying on just one. Comparing the results of one algorithm with those of other algorithms also provides a more complete picture of its performance. In this study, the BTM algorithm stood out as the most suitable for exploratory analysis of social media texts, because it demonstrated a good balance between topic similarity and difference.

4.8 Traditional Algorithms vs. Modern Algorithms

Traditional algorithms (LDA, LSA, PLSA, NMF) consistently showed the highest results for perplexity and the lowest for the other metrics. This reflects their inability to adapt to the textual data used. Considering the results obtained and the characteristics of the texts used in this work, topic modeling carried out by one of these traditional methods may eventually result in sets of words that are relevant to the central topic of the corpus, but it is clear that there is a loss in this product. With higher perplexity and below average coherence (C_{umass}, C_{w2v}), it is clear that these results fall short of what can be achieved by other alternatives.

To illustrate this by a comparison, Table 2 shows examples of topics generated by LDA (traditional) and BTM (co-occurrence) from the same corpus. These topics relate to modeling, where K=10, and three results from each have been selected for display in the table.

As Table 2 shows, the terms present in the LDA topics are quite frequent, especially “covid19”, “coronavirus”, “brasil”, and “morte”. This means that the same words make up almost half of several LDA topics. A crucial reason for this is the frequency of these words in the corpus.

Given the limitations of LDA and other traditional algorithms in dealing with short texts, it is clear that high-

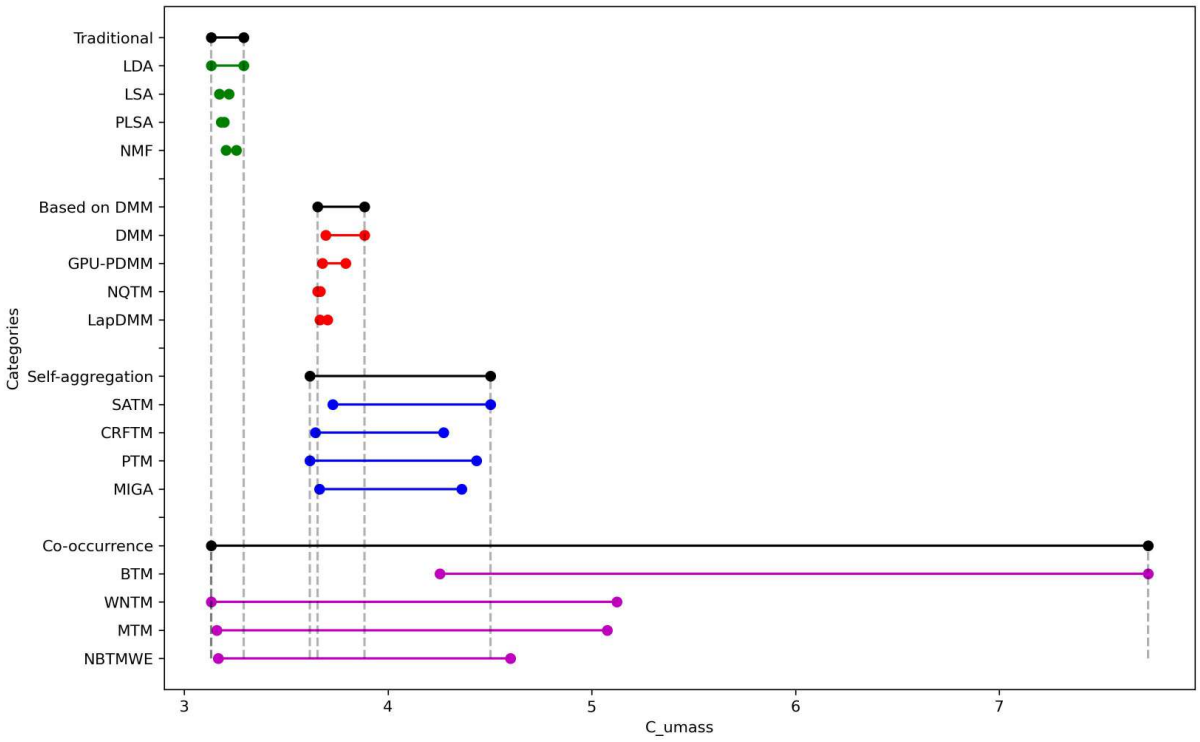

Figure 4. Result ranges for C_{umass}. The colored lines represent the range of values obtained by varying K from 10 to 100.

Table 2. Examples of topics generated by LDA and BTM.

Algorithms	Topic	Words									
LDA	1	covid19	coronavirus	caso	brasil	pessoa	morte	saude	pandemia	casa	morrer
	2	coronavirus	covid19	brasil	caso	pessoa	bolsonaro	morte	presidente	saber	saude
	3	covid19	coronavirus	morte	caso	brasil	bolsonaro	falar	saude	pessoa	morrer
BTM	1	ficar	casa	sair	pessoa	quarentena	medo	bem	longe	doente	mãe
	2	brasil	primeiro	infectado	chegar	italia	morte	número	subir	presidente	nada
	3	vídeo	cardi	falar	coronavirus	pai	mãe	ouvir	preocupar	dia	aula

frequency words were favored in the modeling. For example, the word "coronavirus" appeared in about 30% of the nearly 4 million documents, while "covid19" appeared in about 15% of the texts. It should also be noted that, taken together, the texts in the corpus contained a total of 37,463,729 words.

The BTM, on the other hand, showed a much greater diversity of words and more coherence between them. As can be seen in Table 2, Topic 1 deals with quarantine, people staying at home, and the fear they felt. Topic 2 deals with the first infected person in Brazil and the rising death toll in Italy. The third theme includes the viral video by US artist Cardi B, in which she warmly expresses her fear of the looming COVID-19 outbreak in 2020. Finally, it is important to note that, in this case, due to the high number of documents used and the low K-value, the topics were generated covering multiple themes.

5. Conclusion

The results presented in Section IV indicate the performance of different categories of topic modeling algorithms when applied to short social media texts. The topic distance and symmetric KL-divergence metrics were compared, as both are used to measure the similarity of the generated topics. While these metrics serve similar purposes, they are not exclusive and, when used together, provide a more comprehensive understanding of each algorithm's capabilities.

In general, algorithms based on global co-occurrence showed the lowest perplexity results, followed by self aggregation algorithms. Traditional algorithms, on the other hand, showed the lowest generalization capacity, with the exception (in some cases) of LDA, which showed the most reasonable results in this category. In terms of topic distance and symmetric KL-divergence, the self-aggregation and global co-occurrence algorithms performed best, while the traditional algorithms

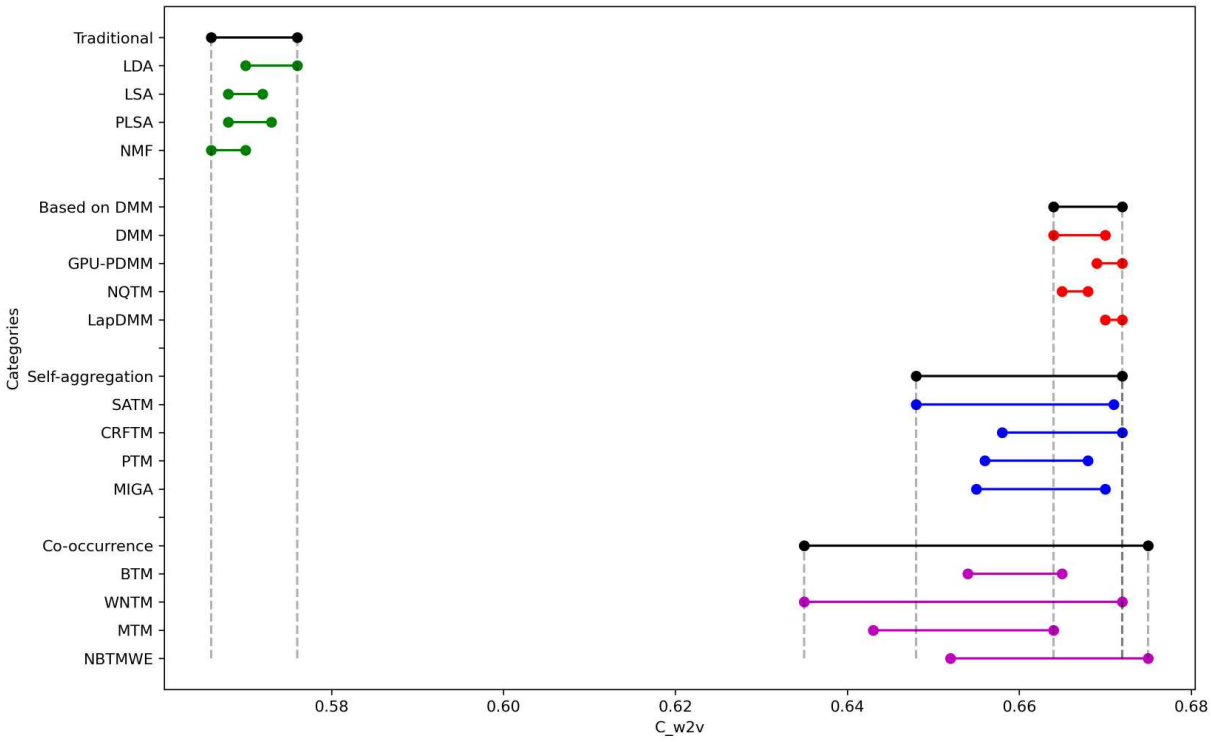

Figure 5. Result ranges for C_{w2v}. The colored lines represent the range of values obtained by varying K from 10 to 100.

showed the lowest values.

For C_{umass} and C_{w2v} coherences, the global co-occurrence algorithms showed the best performance on average, with BTM standing out. The self-aggregation and DMM algorithms followed closely behind, with the traditional algorithms again showing the worst results. In general, algorithms in the same category performed similarly, which suggests that the choice of category-specific technique may be more influential than the choice of a single algorithm.

DMM-based algorithms showed promise for modeling themes from short texts to analyze sub-themes. Although their generalization capacity was lower than that of the self-aggregation and global co-occurrence algorithms, they still stood out for their reasonable results in C_{umass} and C_{w2v} divergence and coherence. On the other hand, the BTM global co-occurrence algorithm stood out for its excellent results in distance, divergence, and C_{umass} coherence, which makes it ideal for exploratory analysis of social media texts.

Finally, the traditional algorithms performed less well in most of the evaluations, with high perplexity and low distances, divergences, and coherences. However, LDA stood out among traditional algorithms, which indicates that, depending on the specific case and the characteristics of the data set, it can be a valid approach to investigate sub-themes in a corpus of texts. The choice of the ideal algorithm therefore depends largely on the objective of the analysis and the specific characteristics of the corpus to be analyzed.

Taken together, these results reinforce the central aim of the study: to provide a comparative perspective on topic modeling approaches. This shows that while global differences exist across algorithmic families, their practical relevance varies by metric, thereby guiding researchers and practitioners in selecting models that best align with the evaluation criteria most critical to their applications.

Acknowledgements

The authors would like to express their gratitude to the Brazilian research agencies that provided support and incentives during the execution of this work, especially Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) and the *Demanda Social* program.

Author contributions

Ian Macedo was the lead author, responsible for the conception of the study, data analysis, and manuscript writing. Luciana Rech and Ricardo Moraes acted as advisors; Luciana provided overall supervision, conceptual guidance, and critical revisions throughout the research process. Ricardo

Moraes contributed substantially to the refinement of the research theme, as well as to the structuring of the theoretical framework and methodological approach.

References

[1] LAUREATE, C. D. P.; BUNTINE, W.; LINGER, H. A systematic review of the use of topic models for short text social media analysis. *Artificial Intelligence Review*, Springer, v. 56, n. 12, p. 14223–14255, 2023.

[2] MCQUILLAN, L. et al. Cultural convergence: Insights into the behavior of misinformation networks on twitter. *arXiv preprint arXiv:2007.03443*, 2020.

[3] NATURE. *The powers and perils of using digital data to understand human behaviour*. 2021. ⟨https://www.nature.com/articles/d41586-021-01736-y/⟩.

[4] BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. *Journal of machine Learning research*, v. 3, n. Jan, p. 993–1022, 2003.

[5] ALASHRI, S. et al. An analysis of sentiments on facebook during the 2016 us presidential election. In: IEEE. *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. [S.l.], 2016. p. 795–802.

[6] ZHANG, Y. et al. idoctor: Personalized and professionalized medical recommendations based on hybrid matrix factorization. *Future Generation Computer Systems*, Elsevier, v. 66, p. 30–35, 2017.

[7] CHEN, Y. et al. What we can do and cannot do with topic modeling: A systematic review. *Communication Methods and Measures*, Taylor & Francis, v. 17, n. 2, p. 111–130, 2023.

[8] WU, X.; LI, C. Short text topic modeling with flexible word patterns. In: IEEE. *2019 International Joint Conference on Neural Networks (IJCNN)*. [S.l.], 2019. p. 1–7.

[9] QIANG, J. et al. Short text topic modeling techniques, applications, and performance: a survey. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, v. 34, n. 3, p. 1427–1445, 2020.

[10] A., A. M. G.; ROBLEDO, S.; ZULUAGA, M. Topic modeling: Perspectives from a literature review. *IEEE Access*, v. 11, p. 4066–4078, 2023.

[11] CHURCHILL, R.; SINGH, L. The evolution of topic modeling. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 54, n. 10s, 2022. ISSN 0360-0300.

[12] YIN, J.; WANG, J. A dirichlet multinomial mixture model-based approach for short text clustering. In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.: s.n.], 2014. p. 233–242.

[13] LI, C. et al. Enhancing topic modeling for short texts with auxiliary word embeddings. *ACM Transactions on Information Systems (TOIS)*, ACM New York, NY, USA, v. 36, n. 2, p. 1–30, 2017.

[14] CHENG, X. et al. Btm: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, v. 26, n. 12, p. 2928–2941, 2014.

[15] HADI, M. A.; FARD, F. H. Aobtm: Adaptive online biterm topic modeling for version sensitive short-texts analysis. In: IEEE. *2020 IEEE international conference on software maintenance and evolution (ICSME)*. [S.l.], 2020. p. 593–604.

[16] GAO, W. et al. Incorporating word embeddings into topic modeling of short text. *Knowledge and Information Systems*, Springer, v. 61, p. 1123–1145, 2019.

[17] ZHAO, H. et al. Leveraging meta information in short text aggregation. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. [S.l.: s.n.], 2019. p. 4042–4049.

[18] FANG, A. Analysing political events on twitter: Topic modelling and user community classification. *SIGIR Forum*, Association for Computing Machinery, New York, NY, USA, v. 53, n. 1, p. 38–39, 2021.

[19] CAMPAGNOLO, J. M.; DUARTE, D.; BIANCO, G. D. Topic coherence metrics: How sensitive are they? *Journal of Information and Data Management*, v. 13, n. 4, 2022.

[20] TOLEGEN, G. et al. A clustering-based approach for topic modeling via word network analysis. In: *2022 7th International Conference on Computer Science and Engineering (UBMK)*. [S.l.: s.n.], 2022. p. 192–197.

[21] WALLACH, H. M. et al. Evaluation methods for topic models. In: *Proceedings of the 26th Annual International Conference on Machine Learning*. [S.l.]: ACM, 2009. (ICML '09), p. 1105–1112.

[22] YUSOF, M. A.; SAEE, S. Code switching: exploring perplexity and coherence metrics for optimizing topic models of historical documents. *International Journal of Systematic Innovation*, v. 8, n. 4, p. 103–118, 2024.

[23] CAO, J. et al. A density-based method for adaptive lda model selection. *Neurocomputing*, v. 72, n. 7, p. 1775–1781, 2009. Advances in Machine Learning and Computational Intelligence.

[24] ARUN, R. et al. On finding the natural number of topics with latent dirichlet allocation: Some observations. In: ZAKI, M. J. et al. (Ed.). *Advances in Knowledge Discovery and Data Mining*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. p. 391–402.

[25] MIMNO, D. et al. Optimizing semantic coherence in topic models. In: *Proceedings of the 2011 conference on empirical methods in natural language processing*. [S.l.: s.n.], 2011. p. 262–272.

[26] LIM, J. P.; LAUW, H. Large-scale correlation analysis of automated metrics for topic models. In: ROGERS, A.; BOYD-GRABER, J.; OKAZAKI, N. (Ed.). *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, 2023. p. 13874–13898.

[27] O'CALLAGHAN, D. et al. An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, Elsevier, v. 42, n. 13, p. 5645–5657, 2015.

[28] BANDA, J. M. et al. A large-scale covid-19 twitter chatter dataset for open scientific research—an international collaboration. *Epidemiologia*, v. 2, n. 3, p. 315–324, 2021. ISSN 2673-3986. Disponível em: ⟨https://www.mdpi.com/2673-3986/2/3/24⟩.