Experimental Evaluation of Topic Modeling Methods for Categorizing Irregularities in Health-related news

Alysson Guimarães[1*], **Methanias Colaço Junior**[1,2], **Samuel Almeida**[1], **Raphael Fontes**[3]
[1]Federal University of Sergipe (UFS), São Cristóvão, SE, Brazil
[2]Laboratory for Technological Innovation in Health (LAIS), UFRN, Natal, RN, Brazil
[3]Center for Innovation and Advanced Technology (NAVI), IFRN, Natal, RN, Brazil
alyssonalk@gmail.com,
{mjrse, samuel16.ti}@hotmail.com,
raphael.fontes@navi.ifrn.edu.br

Abstract

Context: The increasing availability of textual data has driven the application of Natural Language Processing (NLP) techniques in public administration to improve public services. **Objective:** This study aims to analyze topic modeling methods in the context of public health audits conducted by the National Department of SUS Auditing (AudSUS). **Methods:** A controlled in vitro experiment was conducted to assess the performance of the methods in topic modeling tasks using coherence metrics. **Results:** The LSA method stood out among models with the highest average C_V and C_NPMI coherence. LSA-based models achieved superior performance compared to 215 other models in configurations with lower $top - n$ and $top - k$ values. Overall, the statistical analysis confirms that the observed differences among the models are not due to random variation. **Conclusion:** The results underscore the potential of topic modeling methods for clustering news articles that exhibit indications of irregularities, thereby guiding information retrieval during the analytical phase of the audit process. This approach enhances the overall effectiveness of audits and facilitates faster preparation of teams for the operational stage.

1 Introduction

The growing generation of unstructured textual data has fostered the widespread adoption of Natural Language Processing (NLP) techniques within public administration. Governments and institutions increasingly employ these tools to process vast volumes of documents, aiming to enhance the quality of public services, strengthen citizens' trust, and improve efficiency across key sectors such as healthcare, education, and policy decision-making (Jiang et al., 2023).

Within this context, the healthcare sector emerges as particularly vulnerable to fraud and corruption due to the complexity of its systems, the substantial financial resources involved, and the informational asymmetry among stakeholders (Mackey et al., 2018).

Reports indicate that a considerable portion of the global population perceives the healthcare sector as highly prone to corruption, with estimates pointing to financial losses amounting to billions of dollars as a result of fraudulent practices (Mackey et al., 2018). Such misconduct not only causes economic damage but also generates profound social consequences, especially in low-income countries, where it contributes to increased morbidity and mortality rates and deepens inequality. Consequently, mechanisms such as reporting systems and auditing procedures become indispensable, although they encounter significant challenges related to data volume and informational overload (Paula et al., 2024; do Amaral et al., 2020).

The auditing process is inherently costly, time-consuming, and highly dependent on human and material resources, thus demanding the development of solutions capable of automating the analysis of corruption allegations. This process typically unfolds in two stages: the initial identification of potential fraud indicators, followed by a comprehensive investigative phase (Paula et al., 2024).

During the information-gathering phase, diverse sources including websites are utilized (Fontes et al., 2023). To enhance data collection and analytical efficiency, *webscraping* techniques can be integrated with NLP and topic modeling methods for the automated identification the main topics of news content. These approaches substantially reduce the time and costs associated with information processing while supporting the identification of materials that may indicate potential irregularities within the healthcare sector (Benjelloun et al., 2015; Madureira et al., 2021).

Based on this scenario, this article proposes and evaluates the use of topic modeling methods, fol-

*Corresponding author.

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 41–56

lowing the experimental process described in (Colaço Júnior et al., 2022; Colaço Júnior, 2025), to investigate the best methods in terms of coherence metrics and consistency. The main objective is to analyze topic modeling models through a controlled (in vitro) experiment, evaluating them quantitatively with respect to the coherence metrics Coherence Value (C_V) and Normalized Pointwise Mutual Information (C_NPMI). The focus is on the topic modeling of health-related news articles with indications of irregularities, from the perspective of Data Scientists and Auditors of the Brazilian Unified Health System (SUS), in the context of public healthcare audits conducted by the National Department of SUS Auditing (AudSUS).

The article is structured as follows: Section 2 presents a concise literature review on the use of topic modeling for the identification of fraud and corruption, with emphasis on public administration. Section 3 details the dataset employed and the evaluation metrics adopted. In Section 4, the experiment's objective, planning, context selection, research questions, dependent and independent variables, study object selection, experimental design, and instrumentation are described. Section 5 describes the procedures for data preparation, execution, and validation. Subsequently, Section 6 presents the data analysis, interpretation, and statistical evaluation of the results. Section 7 discusses the threats to validity. Finally, Section 8 offers the concluding remarks and outlines potential directions for future research.

2 Related Work

NLP comprises a set of computational approaches designed to evaluate and model natural text across one or more levels of linguistic analysis (Jiang et al., 2023). It is a research field dedicated to investigating and developing methods and systems for the computational processing of human language (Caseli and Nunes, 2024). The field explores how natural human language—both written and spoken—can be processed by computers to perform a variety of tasks. NLP lies at the intersection of multiple disciplines, including linguistics, mathematics, computer science, information science, electrical engineering, artificial intelligence, robotics, and psychology, among others.

Since its emergence in the 1950s, NLP has evolved rapidly, giving rise to a number of classical algorithms and models. Among these is the Term Frequency–Inverse Document Frequency (TF-IDF), which measures the relevance of words within a document. Other widely adopted methods include topic modeling, text classification, and sentiment analysis (Jiang et al., 2023).

Topic Modeling is a fundamental NLP technique (Jiang et al., 2023; Angelov, 2020) and a powerful unsupervised method used to uncover the latent semantic structure within large document collections, typically referred to as topics (Angelov, 2020; Grootendorst, 2022).

The primary goal of topic modeling is to derive concise descriptions of document collections to enable the efficient processing of large text corpora while preserving key statistical relationships relevant to tasks such as classification, novelty detection, summarization, and similarity assessment (Blei et al., 2003). In other words, it aims to automatically discover latent themes (topics) that occur across extensive text datasets. This technique is particularly valuable when the volume of textual data is too large to be manually reviewed or classified (Angelov, 2020).

Topic models are grounded in a probabilistic framework involving topics and documents. A topic represents the subject or theme of a text (Angelov, 2020), and in probabilistic models, it is characterized by a probability distribution over words (Blei et al., 2003, 2006; Teh et al., 2006). Topics are shared across all documents in the corpus (Blei et al., 2006), while each document is viewed as a random mixture of latent topics, with topic proportions being document-specific (Blei et al., 2003; Teh et al., 2006).

This set of techniques has been applied across a wide range of domains, including anti-corruption efforts in public administration.

Driven by the need to address the large volume of textual data and inconsistencies in record entry, resulting from the previously decentralized registration of new items, which hindered price comparison in new procurements and the identification of overpricing, (do Amaral et al., 2020) employed LDA to segment audit data from the Government of Pernambuco (PE) based on the descriptive characteristics of purchased items. The method was applied to a dataset containing 65,000 records of items acquired between 2008 and 2017, with the aim of providing useful information to support oversight activities. The technique proved effective in identifying topics with a finer level of granularity than that achieved through existing human classifi-

cation. Detecting overpricing is one of the primary control activities carried out by the audit team of the State Comptroller General's Office (SCGE) of Pernambuco, serving to prevent acts of administrative misconduct among public officials.

At the height of the COVID-19 pandemic, topic modeling methods were also applied in the context of public health, aiming to identify challenges and opportunities related to vaccination campaigns by analyzing Twitter posts (Jiang et al., 2023).

Historically, topic modeling has evolved from dimensionality reduction techniques such as Latent Semantic Analysis (LSA) (Deerwester et al., 1990), Probabilistic Latent Semantic Indexing (PLSA) (Hofmann, 2013), and Latent Dirichlet Allocation (LDA) (Blei et al., 2003). Over time, several variations and extensions have been proposed, including Non-negative Matrix Factorization (NMF) (Lee and Seung, 1999), Hierarchical Dirichlet Processes (HDP) (Teh et al., 2006), and BERTopic (Grootendorst, 2022). Table 3 in Appendix B describes the evaluated methods in this study.

3 Materials and Methods

This section describes the Database used in this experimentation and the evaluation metrics adopted.

3.1 Database

The database comprises a collection of 154,407 news articles retrieved from the internet. It was developed by (Fontes et al., 2023) through three stages: a proof of concept, an exploratory study, and the final database construction.

To identify health-related news articles containing indications of irregularity, a three-step keyword-based selection strategy was implemented. In the first stage, within the entire database, news articles containing the keyword **"saúde"** in their **content** (corpus) were classified as "Health," while all others were categorized as "Generic News." The generic news articles were excluded from subsequent stages. At this point, 9,516 articles were classified as "Health," and the remaining 144,891 were excluded as generic news.

In the second stage, keywords related to irregularities were applied to the subset of articles previously classified as "Health," dividing them into two groups: those containing irregularity-related keywords and those without, labeled respectively as "Health Irregularity" and "Generic Health." Articles that contained **at least one keyword** in the title and/or abstract (headline) were categorized as "Irregularity News." In this step, 6,239 articles showing indications of irregularity were identified, while the remaining 3,277 were classified as "Generic Health." Table 2 in Appendix A lists the keywords employed.

In the third stage, articles belonging to the "Irregularity News" subgroup were independently assessed by two annotators. A substantial inter-rater agreement (IRA) (Landis and Koch, 1977) between the two evaluators was achieved, with a Cohen's Kappa (k) value of 0.6203. Table 4 in Appendix C presents the contingency table of the evaluations. Five additional evaluators were assigned to resolve cases of disagreement in the classification. Each evaluator was individually responsible for making a final decision in 65 distinct samples. The assessment involved reviewing the titles and abstracts of all articles to confirm their categorization as "Health Irregularity," "Generic Health," or "Generic News."

As a result of the third stage, 421 health-related news articles exhibiting indications of irregularity were identified. Figure 1 illustrates the overall news selection process. The raw and annotated dataset is publicly available on the Zenodo Platform (Guimarães et al., 2025).

3.2 Evaluation Metrics

The following evaluation metrics were used to assess the quality and semantic interpretability of the generated topics: **Coherence Value (C_V)** and **Normalized Pointwise Mutual Information (C_NPMI)** (Röder et al., 2015). These metrics aim to quantify the semantic similarity among the most representative words within each topic, reflecting the degree to which the words tend to co-occur in the reference corpus. Both metrics were implemented using the *CoherenceModel* class from the Gensim library[1].

Coherence Value (C_V) is a metric designed to maximize the correlation with the human evaluation. (C_V) is based on a sliding window over the reference texts. It measures the degree of semantic consistency between the top words of a topic, considering their co-occurrence and contextual relationships. This metric is particularly effective for correlating automated evaluations with human interpretability judgments. The calculation is performed using a sliding window over the reference

[1] https://radimrehurek.com/gensim/

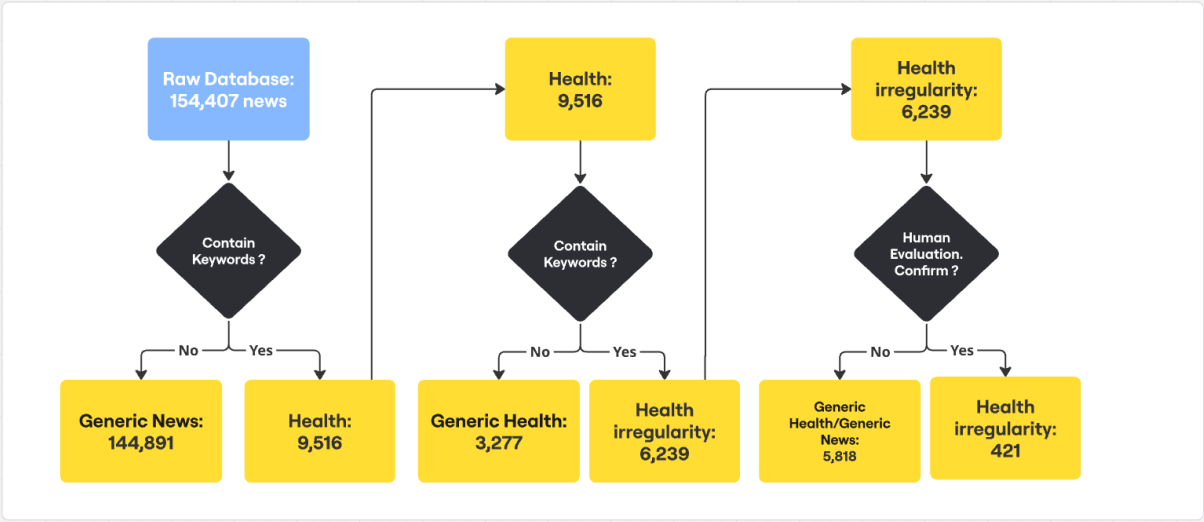

Figure 1: Database classification process.

corpus to estimate co-occurrence probabilities and construct context vectors for each word. The semantic similarity between two words (w_i) and (w_j) is then quantified using the cosine similarity (sim) between their context vectors. The overall coherence of a topic is computed as the mean similarity among all word pairs (P), as shown in Equation 1:

$$\text{C\_V} = \frac{1}{|P|} \sum_{(w_i,w_j)\in P} \text{sim}(w_i, w_j) \qquad (1)$$

Normalized Pointwise Mutual Information (C_NPMI) is derived from the Pointwise Mutual Information (PMI), this metric quantifies the statistical dependence between pairs of words within a topic. The normalization constrains the score to the interval $[-1, 1]$, where higher values indicate stronger positive associations between words, reflecting higher topic coherence, in other words, indicates that the words co-occur more frequently than would be expected by chance, reflecting stronger semantic coherence within the topic. For two words (w_i) and (w_j), with joint probability $(p(w_i, w_j))$ and individual probabilities $(p(w_i))$ and $(p(w_j))$, the NPMI is defined as:

$$\text{NPMI}(w_i, w_j) = \frac{\log \frac{p(w_i,w_j)}{p(w_i)p(w_j)}}{-\log p(w_i, w_j)} \qquad (2)$$

4 Experimental Definition

This section presents the objective of the experimental evaluation, the planning, the research questions, the independent variables, the dependent variables, and the hypotheses.

4.1 Objective

To formalize the objective of this study, the *Goal Question Metric* (GQM) model proposed by (Basili and Weiss, 1984) was adopted. This study aims to **analyze** topic modeling methods through a controlled experiment (*in vitro*), **with the purpose of** evaluating them quantitatively, **with respect to** the metrics of C_V and C_NPMI, **in relation to** the themes (topics) of health-related news with indications of irregularity, **from the perspective of** Data Scientists and Auditors of the Brazilian Unified Health System (SUS), **in the context of** public health audits conducted by the National Department of SUS Auditing (AudSUS).

4.2 Planning

The experiment was conducted in a controlled *in vitro* environment, using the dataset described in Subsection 3.1 and the topic modeling methods described in Section 2. The filtered dataset contains 421 news articles.

4.3 Research Question

To guide the experiment and fulfill the study's objective, the following research questions were formulated:

- RQ1: Which of the selected methods is the best in terms of $coherence$?
- RQ2: Among the selected methods, which are the most consistent in terms of $coherence$?

To address the research questions, the following theoretical hypotheses described in Table 1 were created.

Table 1: Research questions and associated hypotheses

RQ	Null Hypothesis (H_0)	Alternative Hypothesis (H_1)
RQ1	The methods do not have significant differences in terms of coherence.	The methods have significant differences in terms of coherence.
RQ2	The methods are not consistent over the rounds.	The methods are consistent over the rounds.

4.4 Dependent Variables

The dependent variables, or output variables, were the categorized news articles, from which the metrics C_V and C_NPMI can be derived.

4.5 Independent Variables

In this experiment, the independent variables, or input variables, are: the annotated (classified) dataset of news articles with indications of irregularity; and the methods used for the topic modeling task: BERTopic, HDP, LDA, LSA, NMF and pLSA.

4.6 Objects Selection

For the sample size determination, a finite population of 154,407 news articles—representing the total number of items in the complete dataset—was considered. It is important to note that the final sample surpasses the estimated size calculated according to Eq. 4. The sample estimation was based on a 95% confidence level ($Z = 1{,}96$), a tolerable sampling error of 5% ($e = 0{,}05$), and an expected proportion of 50% ($p = 0{,}5$), parameters that maximize sample variability and ensure a conservative estimate of sample size.

The sample size for a finite population was computed in two stages: initially, the sample size for an infinite population (n) was estimated using Eq. 3, followed by an adjustment for a finite population ($n_{adjusted}$) in accordance with Eq. 4, yielding approximately 383.21 samples, as shown in Eq. 6. Ultimately, all manually classified samples from the dataset were employed, 421 classified under "Health Irregularity", thereby exceeding the minimum requirement of 384 articles for a representative sample.

$$n = \frac{Z^2 \cdot p \cdot (1-p)}{e^2} \tag{3}$$

$$n_{adjusted} = \frac{n}{1 + \left(\frac{n-1}{N}\right)} \tag{4}$$

$$n = \frac{1{,}96^2 \cdot 0{,}5 \cdot (1-0{,}5)}{0{,}05^2} = 384{,}16 \tag{5}$$

$$n_{adjusted} = \frac{384{,}16}{1 + \left(\frac{384{,}16-1}{154407}\right)} \approx 383{,}21 \tag{6}$$

4.7 Experiment Design

The execution was designed to follow a systematic process involving preprocessing steps such as lowercasing, stopword removal, stemming (snowball stemming) in brazilian portuguese, implemented with the NLTK python library. In the following, the TF-IDF process aims to capture the relative importance of terms within the news headlines and to generate the dataset's attributes (features) or independent variables described by (Salton and Buckley, 1988) using the Scikit-learn python library. Following the data preparation phase, the topic modeling pipeline was subsequently applied. This preprocessing was not applied to the data processed by BERTopic.

4.8 Instrumentation

The following materials and resources were used:

- Annotated dataset with reference summaries (3.1);
- Python programming language (3.11.13)[2];
- Python libraries: bertopic (0.17.3), gensim (4.3.3), ipykernel (7.0.0), matplotlib (3.10.5), nltk (3.9.2), numpy (2.0.0), openpyxl (3.1.4), pandas (2.3.1), polars (1.34.0), pyarrow (21.0.0), scikit-learn (1.7.1), seaborn (0.13.2), tqdm (4.67.1), and uv (0.8.14);
- Computer with 12th Gen Intel® Core™ i5-1235Ux12 and 16GB RAM;
- Computational resources from the High-Performance Computing Center (NPAD) at the Federal University of Rio Grande do Norte (UFRN).

[2]https://www.python.org/

5 Experiment Operation

This section describes the experiment preparation process, execution, and evaluation of results. Figure 2 describe the experiment preparation and execution performed in the following subsections.

5.1 Experiment Preparation

The preparation of experiment include the creation fo environment with uv library, and creation of the pipelines to generate the results for each topic modeling method. This pipeline comprises a script that executes the procedures for each candidate method described in Subsection 4.5 and for each parameter used, such as top-n (most frequent words) and top-k (number of topics). Six values were considered for both top-n and top-k: 5, 10, 15, 20, 25, and 30. By combining all parameter configurations through a grid search, a total of 216 models (configurations) were generated for the parameters method, top-n, and top-k. Algorithm 1 in the Appendix D presents the details of the execution pipeline. To ensure reproducibility, the random parameter (`random_state`) was assigned to match the current test round, ranging from 1 to 35 for each configuration across the 35 rounds.

As a pilot study, the *pipeline* was initially tested in 5 rounds with 5 samples to verify its operation. Necessary adjustments and potential failures were corrected during this preliminary stage. Subsequently, the full process was executed for all methods.

5.2 Experiment Execution

The experiment was conducted in two stages: the execution of the topic modeling pipeline and results evaluation.

The pipeline illustrated in Figure 2 was executed using Algorithm 1 in the Appendix D. The preprocessing included lowercasing and tf-idf vectorization. All words were converted to lowercase, and subsequently, through the tf-idf transformation.

Subsequently, an evaluation pipeline employing the C_V and C_NPMI metrics was executed, producing the corresponding scores for each topic across every execution round and method, as described in Algorithm 2. Finally, the resulting data were stored for subsequent analysis.

The implementation of both pipelines is publicly available on Github[3].

[3]https://github.com/k3ybladewielder/propor26

5.3 Data Validation

Four (4) statistical tests were employed for data analysis, interpretation, and validation: the Anderson-Darling (AD) test, the Kolmogorov-Smirnov (KS) test, the paired t-test, and the Wilcoxon Signed-Rank test. The Anderson-Darling and Kolmogorov-Smirnov tests were applied to assess the normality of the data. For paired model evaluations exhibiting evidence of normality, the paired t-test was utilized, whereas the Wilcoxon Signed-Rank test was employed to compare the median values of the metrics in cases lacking evidence of normality.

6 Results

This section describes the data analysis and interpretation and the process of statistical evaluation.

6.1 Data Analysis and Interpretation

To address the research questions presented in 4.3, the execution stage was carried out, and the metrics of the topics were obtained for the defined evaluation metrics.

After combining the six methods with the six possible top-n and top-k values, a total of 216 models (configurations) were generated. Each configuration was executed over 35 rounds, and the corresponding results were recorded to compute the mean, median, minimum, maximum, and standard deviation of the coherence metrics. Tables 7 and 8 in Appendix F present the results for the 10 models with the highest average coherence values.

The LSA method stood out compared to the other 215 methods, both in C_V and C_NPMI coherence, followed by NMF and pLSA and LSA and NMF in the first 10 positions for both metrics.

To assess the consistency of the results, the standard deviation was analyzed both collectively and visually across rounds. For comparative purposes, since all metrics range from minus one to one, results were classified based on the standard deviation value for each metric as "Low" ($std < 0.05$), "Moderate" ($std > 0.05$ and < 0.1), and "High" ($std > 0.1$).

From this perspective, among the fifty models with the highest mean C_V values, only three exhibited a moderate standard deviation, while the remaining models presented low values. For the C_NPMI metric, four cases showed moderate standard deviation and forty-six exhibited low variability. A low standard deviation combined with

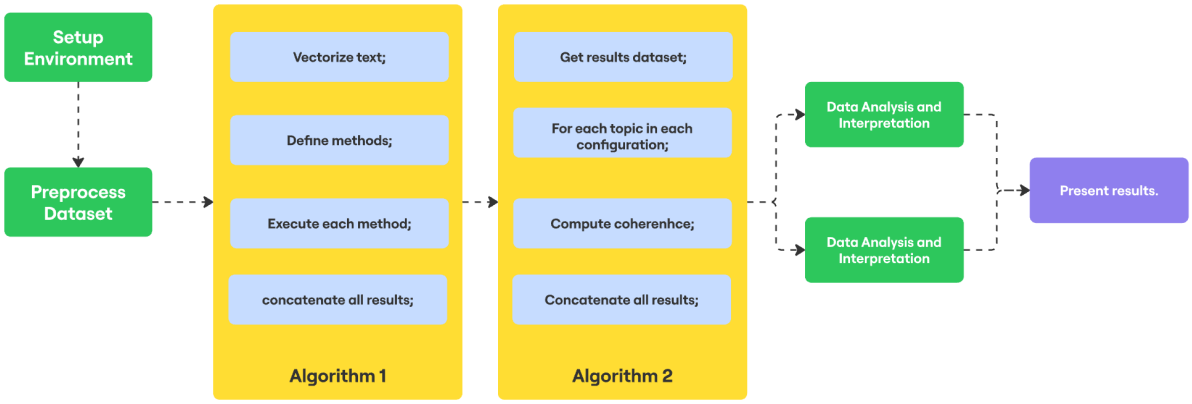

Figure 2: Experiment Preparation and Execution.

high coherence values indicates that the methods are robust to parameter variations. Figures 3 and 4 in Appendix H illustrate these results for the 25 models with the highest C_V and C_NPMI coherence scores.

Among the fifty models with the highest mean C_V values, the LSA method stood out, appearing 20 times, whereas each of the other methods appeared six times. Similarly, among the fifty models with the highest mean C_NPMI values, LSA was again the most frequent, with 19 occurrences, followed by NMF and pLSA with nine each, LDA with seven, and HDP with six.

Regarding the $top-n$ parameter, the best results were obtained using only the five most frequent words, which occurred 36 times. The second-best $top-n$ value was 10 (six occurrences), followed by 15 (five occurrences), while the remaining values appeared only once under the C_V metric. For the C_NPMI metric, the best results also corresponded to top-n = 5 (25 occurrences), followed by $top-n=10$ (17 occurrences), 15 (five occurrences), and the remaining values occurring once each.

Similarly, for the C_V metric, the most effective $top-k$ values were 5 ocurring 11 times, 15, 20, 25, and 30 ocurring 8 times each, and 10 ocurring 7 times. Under the C_NPMI metric, the best $top-k$ values were 15 ocurring 11 times, 20, 25, 30, 5 ocurring 9 times each, and 10 ocurring 5 times, respectively.

6.2 Statistical Evaluation

To compare the relative performance of the algorithms, conclusive statistical evidence is necessary. Accordingly, the Anderson-Darling (AD) and Kolmogorov-Smirnov (KS) tests were applied. The results indicate that certain models and configurations exhibit a normal distribution, as presented in Table 5 and Table 6 in the Appendix E.

The results were divided into two groups: Group A, where there is evidence that the results follow a normal distribution, and Group B, where there is no evidence of normality. To evaluate the performance of paired models in Group A, the paired t-test was used, while the Wilcoxon Signed-Rank Test was applied for Group B. Table 9 describes the number of times the evaluated model outperformed the others.

The results presented in Table 9, which describe the number of cases in which a given model achieved a statistically significantly better result, consistently demonstrate the superior performance of Latent Semantic Analysis (LSA)-based models, particularly in configurations with lower $top-n$ and $top-k$ values. It is observed that LSA variants with $top-n=5$ achieved the highest scores across both coherence metrics, C_V and C_{NPMI}, securing the top positions in the overall ranking. In particular, the configuration with $top-n=5$ and $top-k=5$ emerged as the most effective, outperforming all other 215 evaluated methods, thereby evidencing the robustness and stability of this parametrization.

The observed trend indicates that LSA performance gradually decreases as the $top-k$ parameter increases, suggesting that a smaller set of representative terms is sufficient to maximize topic coherence. This behavior aligns with previous studies highlighting LSA's sensitivity to dimensionality and the selection of relevant terms.

Overall, the statistical analysis confirms that the observed differences among the models are not due to random variation, being statistically significant for both normally distributed groups (tested using the paired t-test) and non-normal groups (evaluated using the Wilcoxon Signed-Rank Test). Therefore, it can be concluded that LSA, particularly in configurations with $top - n = 5$ and $top - k \leq 15$, consistently and statistically outperforms the other methods, establishing itself as the most effective approach for generating coherent topics under the evaluated conditions.

7 Threats to Validity

For the evaluation of the experiment, it is necessary to consider factors that may influence the results, characterized as threats to internal and external validity.

- **Internal Validity**: The topic modeling process was evaluated exclusively through quantitative methods, without human expert assessment. To mitigate this limitation, coherence metrics from the literature—known to exhibit a high correlation with human topic evaluations—were employed.
- **External Validity**: The dataset consists of news articles containing indications of irregularity. Due to the limited linguistic variation among articles, may potentially hinder topic modeling performance. To address this issue, model evaluation was conducted using metrics that have a high correlation with human topic classification.

8 Conclusion and future work

The audit process is typically characterized as costly, time-intensive, and demanding in terms of both human and material resources. In this context, it becomes imperative to implement solutions and techniques that facilitate the automation of corruption complaint analysis.

In this context, with the aim of supporting, enhancing, and optimizing the collection of relevant information that may assist in addressing irregularities, this study presents the results of applying 216 topic modeling configurations to a dataset of health-related news articles exhibiting indications of irregularity. The objective was to assess whether such methods could contribute to the auditing process by directly identifying the underlying themes (topics) of news articles associated with irregularities, as well as determining the optimal set of parameters (configuration) and evaluating their consistency across multiple executions.

Latent Semantic Analysis (LSA)-based models, particularly in configurations with lower $top - n$ and $top - k$ values. It is observed that LSA variants with $top - n = 5$ achieved the highest scores across both coherence metrics, C_V and C_{NPMI}, securing the top positions in the overall ranking. In particular, the configuration with $top - n = 5$ and $top - k = 5$ emerged as the most effective, outperforming all other 215 evaluated methods, thereby evidencing the robustness and stability of this parametrization.

Overall, the statistical analysis confirms that the observed differences among the models are not due to random variation, being statistically significant for both normally distributed groups (tested using the paired t-test) and non-normal groups (evaluated using the Wilcoxon Signed-Rank Test). Therefore, it can be concluded that LSA, particularly in configurations with $top - n = 5$ and $top - k \leq 15$, consistently and statistically outperforms the other methods, establishing itself as the most effective approach for generating coherent topics under the evaluated conditions.

For future work, with the aim of optimizing and automating the auditing process, the following approaches may be explored:

- **Qualitative Evaluation**: In this experiment, the evaluation metrics were exclusively quantitative. In future work, we intend to evaluate the automatically identified topics through qualitative human evaluation.
- **Text Classification**: In this study, a human-labeled dataset was employed. Machine learning models could be used to automatically classify news articles as either containing indications of irregularities or not, thereby automating this time- and labor-intensive classification stage. Such automation would support the prioritization of higher-risk areas during the audit process;
- **Automatic Text Summarization**: The application of text summarization methods, such as language models, can be investigated to reduce information overload while preserving the quality of summaries generated from the large volume of data used in the analytical

phase, thus decreasing the demand for human resources.

9 Limitations

This study was limited to the experimental evaluation of classic methods already established in the literature, not evaluating new modern approaches, such as those based on large language models, with the exception of BERTopic.

References

Dimo Angelov. 2020. Top2vec: Distributed representations of topics. *Preprint*, arXiv:2008.09470.

Victor R. Basili and David M. Weiss. 1984. A methodology for collecting valid software engineering data. *IEEE Transactions on Software Engineering*, SE-10(6):728–738.

Fatima-Zahra Benjelloun, Fatima-Zahra Benjelloun, Ayoub Ait Lahcen, Ayoub Ait Lahcen, Ayoub Ait Lahcen, Samir Belfkih, and Samir Belfkih. 2015. An overview of big data opportunities, applications and tools. *null*.

David Blei, John Lafferty, and 1 others. 2006. Correlated topic models. *Advances in neural information processing systems*, 18:147.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

H. M. Caseli and M. G. V. Nunes, editors. 2024. *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, 2 edition. BPLN.

Methanias Colaço Júnior. 2025. *IA para a Galera Toda: Agentes e Inovação Experimental Sem Código*. Amazon Publishing.

Methanias Colaço Júnior, Rodrigo Cruz, Luciano Araújo, Ana Bliacheriene, and Fátima Nunes. 2022. Evaluation of a process for the experimental development of data mining, ai and data science applications aligned with the strategic planning. *Journal of Information Systems and Technology Management*, 19.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407. First published: September 1990, Citations: 6,569.

João Alberto Arantes do Amaral, João Alberto Amaral, Jairson B. Rodrigues, Jairson Barbosa Rodrigues, and Jairson Barbosa Rodrigues. 2020. Alocação de tópicos latentes - um modelo para segmentação de dados de auditoria do governo de pe. *null*.

Raphael Silva Fontes, Methanias Colaço Júnior, Herrera Prado, Ana Nely, Júlio Araújo, Jailton Carlos de Paiva, and Ricardo Alexsandro de Medeiros Valentim. 2023. Sussurro - detecção na web de eventos auditáveis que representam riscos à saúde pública. *Anais Estendidos do XXIII Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS 2023)*.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Alysson Guimarães, Samuel Almeida, Methanias Colaço Junior, Raphael Fontes, and Gabriely Garcia Ferreira de Araújo. 2025. Health related news dataset.

Thomas Hofmann. 2013. Probabilistic latent semantic analysis. *Preprint*, arXiv:1301.6705.

Yunqing Jiang, Jiaxuan Li, D. Wong, and Ho Yin Kan. 2023. Natural language processing adoption in governments and future research directions: A systematic review. *Applied Sciences*.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Daniel D. Lee and H. Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791.

Tim K. Mackey, Tim K. Mackey, Taryn Vian, Taryn Vian, Jillian Clare Köhler, and Jillian Clare Kohler. 2018. The sustainable development goals as a framework to combat health-sector corruption. *Bulletin of The World Health Organization*.

Luís Madureira, Aleš Popovič, and Mauro Castelli. 2021. Competitive intelligence: A unified view and modular definition. *Technological Forecasting & Social Change*, 173:121086. Received 22 December 2020; Received in revised form 26 July 2021; Accepted 28 July 2021; Available online 9 August 2021; ©2021 Elsevier Inc. All rights reserved.

Thiago De Paula, André Do Amaral, Andre Victor, Luis Alberto Sales, Rodrigo Moreira, Thiago Meirelles, and Rafael Basso. 2024. Automated admissibility of complaints about fraud and corruption. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 610–613, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15)*, pages 399–408, New York, NY, USA. ACM.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2006. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476):1566–1581.

A Keywords

Table 2: Keywords used to identify signs of irregularities.

Keywords
abuso, abuso de poder, acordo ilegal, acusaç, acusação, apropriação indébita, auditoria, aumento orçamento, bilhões, cartel, coação, compra, compras públicas, conluio, contrato, contratos, corrupto, corrupção, corte orçamento, crime, crime organizado, criminoso, deflagrou, denuncia, denúncia, desassistência, desfalque, desonestidade, desonesto, desperdício, desvio, desvios, disfarce, documento alterado, dolo, enganar, engano, enganos, enganoso, enriquecimento, enriquecimento ilícito, escândalo, esquema, evasão, falcatrua, falsa declaração, falsificado, falsificador, falsificação, falso, falta, falta de equipamentos, falta equipamento, fiscalização, forjar, fraudador, fraudar, fraude, fraude em contratos, fraude em licitações, fraude financeira, fraude licitação, fraudulento, fugir, golpe, ilegal, ilusão, ilícito, indicativo, indício, infração, investiga, investigação, irregular, irregularidade, irregularidade administrativa, irregularidade de gestão, irregularidade financeira, irregularidades, lavagem, lavagem de dinheiro, licitação, mandado, manipulado, manipulador, manipulação, manipulação de dados, maquiar, mentira, milhões, má conduta, negligência dolosa, ocultar, ocultação, PF, peculato, perjúrio, plano, plano de saúde, plano saúde, prevaricação, propina, recurso, relatório falso, rombo, roubo, sem autorização, sem consentimento, sobrefaturamento, sonegar, sonegação, suborno, sugestão, superfaturamento, suspeit, suspeita, suspeito, transação, transação suspeita, transgressão, transparência, uso indevido, uso indevido de recursos, plano saude, uso irregular, venda.

B Topic Modeling Algorithms

Table 3: Description of topic modeling algorithms.

Algorithm	Description
Latent Semantic Analysis (LSA)	LSA, also known as Latent Semantic Indexing (LSI), is a method for automatic indexing and retrieval that seeks to exploit the higher-order structure implicit in the association of terms with documents to enhance the detection of relevant materials. This approach addresses issues related to synonymy and polysemy by modeling latent semantic structures hidden in the term–document matrix. To estimate this structure, LSA applies Singular Value Decomposition (SVD), which captures co-occurrence patterns among words and documents (Deerwester et al., 1990).
Probabilistic Latent Semantic Indexing (pLSA)	pLSA, or the Aspect Model, is a probabilistic extension of LSA designed to model co-occurrence data statistically. Instead of using purely algebraic decomposition, pLSA employs a latent class model to represent documents as mixtures of topics, each topic being a probability distribution over words. This statistical foundation enables pLSA to better capture polysemy and word–context dependencies (Hofmann, 2013).
Latent Dirichlet Allocation (LDA)	LDA is a generative probabilistic model that represents documents as mixtures over latent topics, where each topic is characterized by a distribution over words. It introduces Dirichlet priors to control topic–document and word–topic distributions, allowing more interpretable topic discovery. LDA provides a Bayesian framework for uncovering the latent thematic structure in large text corpora (Blei et al., 2003).
Non-Negative Matrix Factorization (NMF)	NMF is a matrix factorization algorithm that decomposes a non-negative data matrix—such as a term–document matrix—into two lower-dimensional non-negative matrices, enabling part-based and additive representations. The method learns semantic components (W) and encodings (H) simultaneously, producing interpretable latent structures that align with human-understandable "topics" in textual data (Lee and Seung, 1999).
Hierarchical Dirichlet Processes (HDP)	HDP is a Bayesian nonparametric model that extends the Dirichlet Process (DP) to handle grouped data. It enables automatic inference of the number of clusters (topics) without pre-specifying this number. HDP models each group as a mixture of an unknown number of components, which can be shared across groups, allowing greater flexibility and generalization in topic modeling (Teh et al., 2006).
BERTopic	BERTopic is a modern topic modeling method that leverages Transformer-based text embeddings to detect semantically coherent clusters of documents. It uses the representational power of language models (e.g., SBERT), dimensionality reduction (e.g., UMAP), and density-based clustering (e.g., HDBSCAN), followed by class-based TF-IDF to extract interpretable topic representations. Unlike traditional bag-of-words models, BERTopic preserves semantic relationships among words (Grootendorst, 2022).

C Contingency Matrix

Table 4: Contingency matrix of annotations between Evaluator 1 and Evaluator 2.

Evaluator 1 \Evaluator 2	**Generic News**	**Health (Generic)**	**Health (irregularity)**	**Total Linha (R_i)**
Generic News	4532	302	145	4979
Health (generic)	146	676	7	829
Health (irregularity)	211	17	203	431
Total (C_i)	**4889**	**995**	**355**	$N = 6239$

D Algorithms

Algorithm 1 Topic Modeling Experimentation Pipeline

Input: Dataset df, column name $text\_col$, number of topics n_topics, number of rounds n_rounds
Extract texts: $texts \leftarrow df[text\_col]$ (fill missing, convert to string)
Initialize **vectorizer** $\leftarrow$ TF-IDF(max_features = 5000)
Compute document-term matrix $X \leftarrow vectorizer.fit\_transform(texts)$
Initialize empty list $results \leftarrow []$
Define **methods** $\leftarrow$ [NMF, LSA, LDA, BERTopic, HDP, pLSA]
for each **method** in methods **do**
  for each round $i = 1$ **to** n_rounds **do**
    Train $method(texts)$
    $labels \leftarrow topics$
    Create temporary dataframe $df\_tmp \leftarrow df.copy()$
    Add columns: method, round = i, topic = $labels$
    Append df_tmp to $results$ exception e
    Print error message $[ERRO]\ method, round, e$
  end for
end for
Concatenate all partial results $df\_final \leftarrow concat(results)$
Return df_final

Algorithm 2 Topic Coherence Evaluation Function

Input: Dataset df_topics, columns $(text\_col, topic\_col, method\_col, round\_col)$, metrics set $metrics = \{c\_v, c\_npmi\}$, and top-$n$ words $topn$
Tokenize texts: $df\_topics[\_tokens\_] \leftarrow split(df\_topics[text\_col], "")$
Initialize empty list $results \leftarrow []$
Partition dataset by $(method\_col, round\_col)$ into groups $grouped$
for each $configuration$ **do**
    Extract tokenized documents: $tokenized\_docs \leftarrow group\_df[\_tokens\_]$
    Build dictionary: $dictionary \leftarrow Dictionary(tokenized\_docs)$
    Build corpus: $corpus \leftarrow [dictionary.doc2bow(doc)$ for each $doc \in tokenized\_docs]$
    Initialize empty list $topics \leftarrow []$
    for each $metric$ in $metrics$ **do**
        Compute coherence model
        $cm \leftarrow CoherenceModel(topics, tokenized\_docs, dictionary, metric)$
        Get coherence score $coh \leftarrow cm.get\_coherence()$
        Append result $\{method, rnd, metric, coh, |topics|, |group\_df|, top - n\}$ to $results$
    end for
    exception e
    Print error message $[ERROR]$ $method, rnd, e$
end for
Return $DataFrame(results)$

E Normality Test

Table 5: Normality test results (Anderson–Darling and Kolmogorov–Smirnov) for C_V metric across models and configurations. Critical value of 0.719 and N = 35. Showing only the 25 with the highest average coherence, but applied to all 216 models.

Config	Metric	AD_Statistic	AD_Critical_5%	AD_Reject_at_5%	KS_Statistic	KS_pvalue
model: HDP, top-n: 5, top-k: 10	C_V	0.284013	0.719	False	0.085657	9.401347e-01
model: HDP, top-n: 5, top-k: 15	C_V	0.692053	0.719	False	0.138716	4.693841e-01
model: HDP, top-n: 5, top-k: 20	C_V	0.300185	0.719	False	0.097755	8.594782e-01
model: HDP, top-n: 5, top-k: 25	C_V	0.549617	0.719	False	0.139753	4.600110e-01
model: HDP, top-n: 5, top-k: 30	C_V	0.202301	0.719	False	0.066194	9.952686e-01
model: HDP, top-n: 5, top-k: 5	C_V	0.350576	0.719	False	0.083276	9.517420e-01
model: LDA, top-n: 5, top-k: 15	C_V	0.623572	0.719	False	0.138019	4.757282e-01
model: LDA, top-n: 5, top-k: 20	C_V	0.221110	0.719	False	0.081094	9.610818e-01
model: LDA, top-n: 5, top-k: 25	C_V	0.263966	0.719	False	0.099633	8.441219e-01
model: LDA, top-n: 5, top-k: 30	C_V	0.550680	0.719	False	0.106294	7.851221e-01
model: LSA, top-n: 5, top-k: 10	C_V	0.894590	0.719	True	0.168011	2.473601e-01
model: LSA, top-n: 5, top-k: 15	C_V	0.509020	0.719	False	0.107231	7.763647e-01
model: LSA, top-n: 5, top-k: 20	C_V	0.324642	0.719	False	0.115306	6.978067e-01
model: LSA, top-n: 5, top-k: 25	C_V	0.431313	0.719	False	0.128323	5.679148e-01
model: LSA, top-n: 5, top-k: 30	C_V	0.410984	0.719	False	0.105012	7.969562e-01
model: LSA, top-n: 5, top-k: 5	C_V	8.971564	0.719	True	0.476264	8.136301e-08
model: NMF, top-n: 5, top-k: 10	C_V	5.425697	0.719	True	0.396011	1.840425e-05
model: NMF, top-n: 5, top-k: 15	C_V	5.947324	0.719	True	0.412082	6.834794e-06
model: NMF, top-n: 5, top-k: 20	C_V	2.130595	0.719	True	0.216880	6.331921e-02
model: NMF, top-n: 5, top-k: 25	C_V	0.653413	0.719	False	0.121690	6.338602e-01
model: NMF, top-n: 5, top-k: 30	C_V	0.195797	0.719	False	0.076759	9.760541e-01
model: pLSA, top-n: 5, top-k: 15	C_V	1.900900	0.719	True	0.224083	5.026736e-02
model: pLSA, top-n: 5, top-k: 20	C_V	0.611442	0.719	False	0.110636	7.437987e-01
model: pLSA, top-n: 5, top-k: 25	C_V	0.478568	0.719	False	0.124746	6.033013e-01
model: pLSA, top-n: 5, top-k: 30	C_V	0.291882	0.719	False	0.073135	9.851288e-01

Table 6: Normality test results (Anderson–Darling and Kolmogorov–Smirnov) for C_NPMI metric across models and configurations. Critical value of 0.719 and N = 35.

Config	Metric	AD_Statistic	AD_Critical_5%	AD_Reject_at_5%	KS_Statistic	KS_pvalue
model: LDA, top-n: 5, top-k: 15	C_NPMI	1.083289	0.719	True	0.150547	0.368681
model: LDA, top-n: 5, top-k: 20	C_NPMI	0.324173	0.719	False	0.095154	0.879625
model: LDA, top-n: 5, top-k: 25	C_NPMI	0.163557	0.719	False	0.067908	0.993511
model: LDA, top-n: 5, top-k: 30	C_NPMI	0.539028	0.719	False	0.125273	0.598055
model: LSA, top-n: 10, top-k: 10	C_NPMI	0.618090	0.719	False	0.128642	0.564785
model: LSA, top-n: 10, top-k: 15	C_NPMI	2.326013	0.719	True	0.230684	0.040407
model: LSA, top-n: 10, top-k: 20	C_NPMI	0.280068	0.719	False	0.101739	0.826180
model: LSA, top-n: 10, top-k: 25	C_NPMI	0.214404	0.719	False	0.095555	0.876611
model: LSA, top-n: 10, top-k: 30	C_NPMI	0.444250	0.719	False	0.117572	0.675169
model: LSA, top-n: 10, top-k: 5	C_NPMI	5.658704	0.719	True	0.433285	0.000002
model: LSA, top-n: 15, top-k: 5	C_NPMI	9.476502	0.719	True	0.438859	0.000001
model: LSA, top-n: 5, top-k: 10	C_NPMI	0.970999	0.719	True	0.161610	0.287937
model: LSA, top-n: 5, top-k: 15	C_NPMI	0.547248	0.719	False	0.127862	0.572444
model: LSA, top-n: 5, top-k: 20	C_NPMI	0.461266	0.719	False	0.108281	0.766434
model: LSA, top-n: 5, top-k: 25	C_NPMI	0.309672	0.719	False	0.100200	0.839363
model: LSA, top-n: 5, top-k: 30	C_NPMI	0.310623	0.719	False	0.099127	0.848324
model: LSA, top-n: 5, top-k: 5	C_NPMI	7.059189	0.719	True	0.405031	0.000011
model: NMF, top-n: 5, top-k: 15	C_NPMI	4.997353	0.719	True	0.378835	0.000050
model: NMF, top-n: 5, top-k: 20	C_NPMI	3.069506	0.719	True	0.275812	0.007611
model: NMF, top-n: 5, top-k: 25	C_NPMI	0.576560	0.719	False	0.138278	0.473371
model: NMF, top-n: 5, top-k: 30	C_NPMI	0.299215	0.719	False	0.113968	0.711104
model: pLSA, top-n: 5, top-k: 15	C_NPMI	0.935512	0.719	True	0.200904	0.102791
model: pLSA, top-n: 5, top-k: 20	C_NPMI	0.473631	0.719	False	0.099061	0.848867
model: pLSA, top-n: 5, top-k: 25	C_NPMI	0.293222	0.719	False	0.085942	0.938645
model: pLSA, top-n: 5, top-k: 30	C_NPMI	0.186404	0.719	False	0.083504	0.950693

F Coherence Metrics

Table 7: Coherence statistics for the best 10 topic models using the C_V metric. Sorted by mean coherence.

Metric	Config	Mean	Median	Min	Max	Std	Class
C_V	model: LSA, top-n: 5, top-k: 5	0.678336	0.667915	0.667915	0.719436	0.020852	Low
C_V	model: LSA, top-n: 5, top-k: 15	0.659714	0.659786	0.647247	0.681399	0.008474	Low
C_V	model: LSA, top-n: 5, top-k: 20	0.649771	0.650636	0.627398	0.671620	0.010073	Low
C_V	model: LSA, top-n: 5, top-k: 10	0.642928	0.643777	0.588675	0.705938	0.028154	Low
C_V	model: LSA, top-n: 5, top-k: 25	0.641559	0.644809	0.611926	0.662006	0.012006	Low
C_V	model: LSA, top-n: 5, top-k: 30	0.637768	0.637706	0.614423	0.656624	0.010692	Low
C_V	model: NMF, top-n: 5, top-k: 25	0.628399	0.629427	0.615936	0.639792	0.007176	Low
C_V	model: NMF, top-n: 5, top-k: 30	0.627461	0.627654	0.613681	0.637752	0.005258	Low
C_V	model: pLSA, top-n: 5, top-k: 30	0.623795	0.624451	0.599529	0.639714	0.010080	Low
C_V	model: pLSA, top-n: 5, top-k: 25	0.618417	0.616701	0.604708	0.635379	0.007992	Low

Table 8: Coherence statistics for the best 10 topic models using the C_NPMI metric. Sorted by mean coherence.

Metric	Config	Mean	Median	Min	Max	Std	Class
C_NPMI	model: LSA, top-n: 5, top-k: 5	0.175260	0.158527	0.158527	0.231728	0.029230	Low
C_NPMI	model: LSA, top-n: 5, top-k: 15	0.154070	0.153015	0.134467	0.183592	0.012860	Low
C_NPMI	model: LSA, top-n: 5, top-k: 10	0.142514	0.143693	0.086927	0.217466	0.037734	Low
C_NPMI	model: LSA, top-n: 5, top-k: 20	0.141044	0.143768	0.116590	0.165643	0.012833	Low
C_NPMI	model: LSA, top-n: 5, top-k: 25	0.130320	0.133465	0.092675	0.156592	0.014876	Low
C_NPMI	model: LSA, top-n: 5, top-k: 30	0.128705	0.129010	0.083141	0.155468	0.015897	Low
C_NPMI	model: LDA, top-n: 5, top-k: 30	0.115689	0.107268	0.049664	0.201930	0.040176	Low
C_NPMI	model: LDA, top-n: 5, top-k: 25	0.112683	0.113851	-0.023658	0.236730	0.054868	Moderate
C_NPMI	model: LSA, top-n: 10, top-k: 5	0.110963	0.111709	0.089543	0.117318	0.006333	Low
C_NPMI	model: NMF, top-n: 5, top-k: 30	0.108545	0.108976	0.090646	0.121334	0.007590	Low

G Model Performance

Table 9: Total scores of the 25 best models by coherence metric (C_NPMI and C_V) and overall sum.

Model	C_NPMI	C_V	Total
model: LSA, top-n: 5, top-k: 5	215	215	430
model: LSA, top-n: 5, top-k: 15	213	214	427
model: LSA, top-n: 5, top-k: 20	212	212	424
model: LSA, top-n: 5, top-k: 10	212	210	422
model: LSA, top-n: 5, top-k: 25	210	210	420
model: LSA, top-n: 5, top-k: 30	209	210	419
model: NMF, top-n: 5, top-k: 25	203	208	411
model: NMF, top-n: 5, top-k: 30	203	208	411
model: pLSA, top-n: 5, top-k: 30	201	206	407
model: LDA, top-n: 5, top-k: 30	204	199	403
model: LDA, top-n: 5, top-k: 25	203	199	402
model: NMF, top-n: 5, top-k: 20	194	201	395
model: pLSA, top-n: 5, top-k: 25	194	201	395
model: pLSA, top-n: 5, top-k: 15	194	200	394
model: pLSA, top-n: 5, top-k: 20	193	201	394
model: LDA, top-n: 5, top-k: 20	196	197	393
model: NMF, top-n: 5, top-k: 15	193	199	392
model: LSA, top-n: 10, top-k: 5	204	186	390
model: LSA, top-n: 10, top-k: 15	202	186	388
model: LSA, top-n: 10, top-k: 20	194	180	374
model: LSA, top-n: 10, top-k: 10	194	174	368
model: LSA, top-n: 15, top-k: 5	196	169	365
model: LSA, top-n: 10, top-k: 25	189	175	364
model: HDP, top-n: 5, top-k: 10	171	190	361
model: LSA, top-n: 10, top-k: 30	186	175	361

H Boxplots

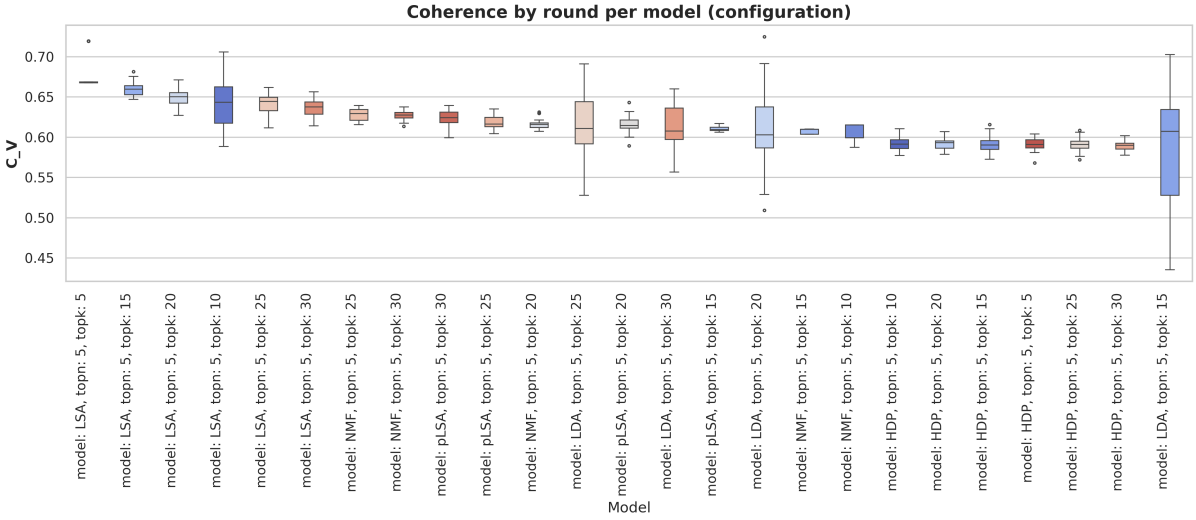

Figure 3: Boxplot of the 25 models with highist mean C_V coherence.

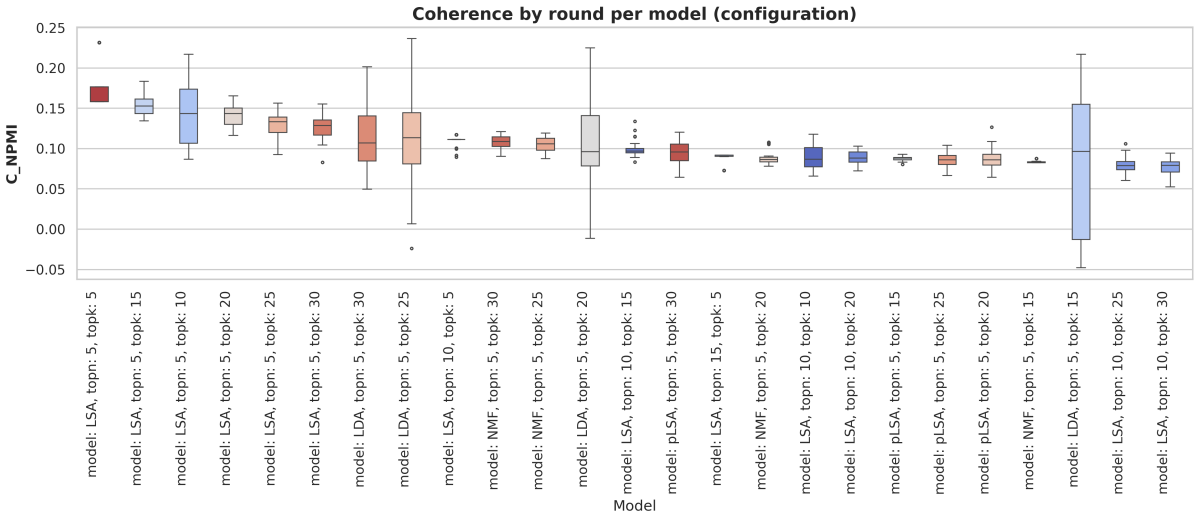

Figure 4: Boxplot of the 25 models with highist mean C_NPMI coherence.