# From LDA to BERTopic: Evaluating Topic Modelling Methods for Aviation Safety Reports in Brazilian Portuguese

MSc Research Project
Data Analytics

Priscila Cristina da Silva de Oliveira
X23157003

School of Computing
National College of Ireland

Supervisor: Barry Haycock

**National College of Ireland**

**MSc Project Submission Sheet**

**School of Computing**

**Student Name:** Priscila Cristina da Silva de Oliveira

**Student ID:** X23157003

**Programme:** MSc in Data Analytics **Year:** 2024

**Module:** Research Project

**Supervisor:** Barry Haycock

**Submission Due Date:** 16.09.2024

**Project Title:** From LDA to BERTopic: Evaluating Topic Modelling Methods for Aviation Safety Reports in Brazilian Portuguese

**Word Count:** 5632 **Page Count** 19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

**Signature:** Priscila C. D. Oliveira

……………………………………………………………………………………………………………

**Date:** 16.09.2024

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST**

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies) | □ |
| **Attach a Moodle submission receipt of the online project submission,** to each project (including multiple copies). | □ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | □ |

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# From LDA to BERTopic: Evaluating Topic Modelling for Aviation Safety Reports in Brazilian Portuguese

Priscila Cristina da Silva de Oliveira

X23157003

**Abstract**

This study applies five different topic models to aviation safety reports in Brazilian Portuguese. The techniques explored are Latent Dirichlet Allocation (LDA), LDA with stemming, a cross-language model which translates the texts to English and then perform LDA, word2vec with k-means and BERTopic. The research aims to explore the dataset that was not previously used in published research and evaluate how effective the approaches applied are in identifying topics withing the corpus of reports. BERTopic outperformed the other models achieving a coherence score of 0.4819. A composite score was calculated based on the coherence and perplexity scores and used to evaluate the LDA models. LDA with stemming demonstrated the best composite score. Furthermore, Word2Vec with k-means might be a better approach for more generalised classifications.

## 1 Introduction

Just three decades ago, travelling by air was a luxury reserved for the most fortunate, often viewed as a status symbol. Nowadays, air travel has become increasingly more accessible and, consequently, a common mode of transportation for the average person. The increased demand in air travel has led to a higher frequency of flights, making aviation safety more important than ever. Analysing aviation safety reports is an insightful way to learn from the events that happened in the past and avoid similar ones in the future. The large volume of documents presents a significant challenge for those responsible for the analysis of the aviation safety reports, and this is where the use of topic modelling becomes particularly useful.

Topic modelling is a technique used in natural language processing (NLP) and machine learning (ML) to uncover latent topics in a corpus of texts by identifying sets of words and/or phrases which best represent each topic (Alammar and Grootendorst, 2024). The method can bring to light issues that were not previously unveiled. In spite of that, applying topic modelling to non-English languages can demonstrate disparities in relation to the number of specialised NLP tools available for different languages (Baishya and Baruah, 2022) which is no different for Brazilian Portuguese.

There are two important observations to make when working with topic models. Firstly, the technique does not determine the number of topics within the corpus. The number is defined by the analyst working on the model. It is a process of trial and error until finding the best number of topics. Secondly, it does not name the topics either, but returns a series of words which represent each topic. For this reason, a certain level of expertise in the domain being analysed is necessary in order to classify the topics found.

This study aims to evaluate and compare how effective different topic modelling approaches are when applied to aviation safety reports in Brazilian Portuguese. Specifically, five different models are analysed: Latent Dirichlet Allocation (LDA), LDA with stemming, an English translation approach followed by LDA, Word2Vec combined with K-means clustering, and, finally, BERTopic.

The remainder of this paper is structured as follows: Section 2 provides a review of relevant literature. Section 3 details the methodology used, including data collection, preprocessing steps, and implementation of each model. Section 4 presents the design specification. Section 5 give details of the implementation. Section 6 provides the evaluation conducted. Finally, Section 7 the study is concluded.

# 2 Related Work

LDA, created by David M. Blei, Andrew Y. Ng and Michael I. Jordan in 2002, was the first topic model. Since its release, many more sophisticated methods to uncover latent topics in a corpus of texts have been introduced. The new approaches have incorporated NLP techniques such as word embeddings and neural networks (Churchill and Singh, 2021). The characteristics of the data to be analysed is something to consider when selecting a topic modelling technique. The volume and length of the documents, for example, can affect the performance of the method used (Churchill and Singh, 2021).

Various NLP techniques have been recently explored to analyse aviation safety reports. Buselli *et al.* (2022) applied a combination of LDA, clustering and syntactic analysis to extract meaningful information from safety reports, focusing on loss of separation. The authors obtained promising results identifying main topics in the corpus analysed and grouping incidents which revealed patterns such as pilot errors and air traffic controller's coordination issues. They highlighted that the success of the analysis relies on the quality of the reports and the lack of standardisation can pose a challenge. Tikayat Ray *et al.* (2023) went further and explored generative language models to produce synopses of the occurrences, identify human factors, and attribute responsibility. They have also applied embeddings from BERT, aeroBERT, and sBERT to fine-tune and improve the performance of their model. Although the study has demonstrated the high potential of the approaches chosen, it does not come without caution. Aviation safety reports are not to be used for purposes other than preventing future occurrences from happening. From the moment document analysis shift to assigning blame for the incidents, the main purpose of the analysis is lost, as it deviated from its main purpose.

These are a few examples of how researchers have been applying NLP to study aviation safety reports. This literature review provides a comprehensive overview of the relevant research. It focuses on LDA, Word2vec, and BERTopic, which are the approaches employed in this study. It also explores the use of NLP tools for the Portuguese language.

## 2.1 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation, commonly known by LDA, is still heavily used 20 years after its creation. Kim *et al.* (2023) analysed Korean aviation safety reports by performing LDA. The authors achieved 75% accuracy, when comparing the topics obtained with those defined by experts. Their study emphasises the importance of pre-processing the data: removing special characters and stop words. They used the coherence and perplexity scores to determine the optimal number of topics. One point of attention is the size of the dataset used by the authors, which consists of forty documents only, which might have affected the performance of the model, as stated by Churchill and Singh (2021). Luo and Shi (2019) proposed using a different version of the LDA model to analyse aviation safety reports. The lda2vec technique combines elements of LDA and word2vec word embeddings resulting in better capturing local and global semantic information. Similarly to the previous study, the authors relied on the coherence score to determine the optimal number of topics. When compared to LDA, the lda2vec model achieved higher values of coherence score.

Kuhn (2018) explored Structural Topic Model (STM), another variation of the traditional LDA. STM incorporates metadata. The author used the metadata to examine how the frequency of topics changes with the phases of the flight, mission type, and time. The study identified several topics in the data analysed and found cyclical patterns in issues related to air traffic control. Xing *et al.* (2024) identified topics in aviation safety reports using similar approach (LDA and STM models), although employing different datasets. Their study used TF-IDF for text vectorisation and built a word co-occurrence network to help visualise relationships between the identified risk factors. The methods were compared for validation, with STM outperforming LDA. Not only has it provided more granular topics, but it has also better distinguished similar occurrences. An interesting finding was that weather-related topics were more prevalent in fatal events. Common challenges such as dealing with aviation-specific jargons and multi word expressions were mentioned by the authors. This demonstrates once again the importance of the standardisation of the language used when reporting aviation occurrences.

All authors have somehow emphasised the importance of domain expertise to evaluate the performance of the models, with Kim *et al.* (2023) relying on experts to manually label the topics.

## 2.2 Word2vec

Combinations of word2vec, a method for converting texts into word vectors, and k-means, an unsupervised clustering algorithm, have been explored as an alternative to traditional topic models and text summarisation. An example is the work of De Miranda, Pasti and De Castro (2020) which used word2vec to generate distributed word representations and detected topics in a corpus of emails by clustering the documents using Self-Organizing Maps (SOM) and k-means. Even though the approach effectively identified different topics, similar themes would sometimes be considered a single topic, demonstrating that the approach might not be the best option in situations where a certain level of detail is required. Haider *et al.* (2020) leveraged the combination of word2vec with k-means to perform automatic text summarisation by introducing a sentence-based clustering approach. The authors did not directly explore topic modelling, but they have demonstrated the potential of the technique to extract relevant topics from documents.

## 2.3 BERTopic

BERTopic is a topic modelling approach recently created by Grootendorst (2022). It uses pre-trained language models, clustering, UMAP for dimensionality reduction, and TF-IDF for text vectorisation. One limitation of the technique is that BERTopic assumes each document contains only one topic, what does not always reflect reality. The model often outperforms traditional techniques such as LDA but it has the drawback of requiring higher computation time. Although not vastly used to analyse aviation safety reports, it has been employed in different domains as the topic modelling approach. Maschek and Stöckl (2023), for example, used the technique to automate the analysis of sport and leisure accident reports, by clustering preventive measures into related topics. The authors found that the topic clusters created by BERTopic were more interpretable when compared to LDA results, demonstrating the potential of the technique for the analysis of reports.

## 2.4 NLP Tools and Languages Other than English

Despite the fact that NLP techniques have been largely used to analyse and obtain insights from aviation safety reports, exploring the techniques for similar documents in other languages is a research topic that has not been fully explored. However, one should not take this research gap as a sign of lack of interest of the non-English speakers working on the field. The disparity on the resources available for the development of new NLP tools results in "underrepresentation of non-English languages in NLP research and models" as well stated by Wang (2023). Efforts have been made to develop new tools and adapt the existing ones. An example is the Python pipeline NLPyPort described by Ferreira, Gonçalo Oliveira and Rodrigues (2019) which process documents in Portuguese.

# 3 Research Methodology

The Brazilian dataset of aviation safety reports has not yet been studied in previously published research papers. This presents opportunities for further research to be conducted, providing deeper understanding of the documents and of the performance of topic models applied to aviation-related documents in a language other than English. Five models were proposed. The traditional LDA method will be compared to two other approaches using the same technique: an LDA model with stemming and a cross-language model which translates the documents to English and then performs LDA. The other two techniques used are a combination of word2vec followed by k-means and BERTopic. The objective is to analyse and compare the performance of the different approaches when applied to documents written in Brazilian Portuguese.

The Knowledge Discovery in Databases (KDD) process proposed by Fayyad, Piatetsky-Shapiro and Smyth (1996) is the framework used in this study. Figure 1 illustrates the main steps of the process.

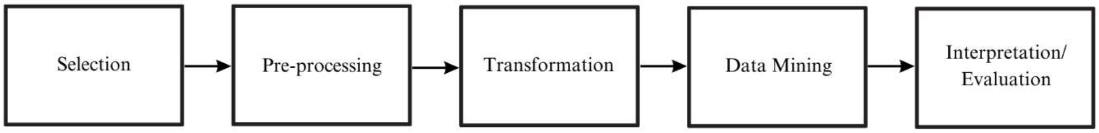

**Figure 1: These are the five main steps of the methodology used in this study given by KDD.**

## 3.1 Data Selection

Brazilian aviation safety reports are made publicly available by the Brazilian regulatory body CENIPA[1] (Aeronautical Accidents Investigation and Prevention Centre). The dataset available consists of 3659 aviation occurrences classified in 32 categories (topics). Each occurrence is documented by at least one report, typically in Portuguese and, in some cases, in a second language. This research will focus exclusively on the 3099 Portuguese, machine-readable files which were stored on the local machine.

## 3.2 Data Pre-processing and Transformation

The data pre-processing and transformation phase consisted of the steps given by Rose *et al.* (2022). The texts were extracted, cleaned and, in some cases, normalised before the implementation of the models. The raw texts were initially extracted from the PDF files. In the cleaning phase, texts were converted to lower case, numbers and punctuation were removed. Next, the raw texts were tokenised, split into individual words, and, finally, the stop words removed. The list of stop words utilised was customised by adding aviation-specific terms, and other vocabulary likely to be in the reports but with no valuable meaning for the analysis.

Stemming was applied to one of the LDA models with the objective to analyse if and how the process would affect the topics obtained. Regarding the cross-language model, the translation of the documents was performed as the first step of the pre-processing stage. The translation process has demonstrated to be the most time-intensive aspect of the pre-processing and transformation phases. To optimise the process, once translated the texts were saved in a pickle file.

[1] CENIPA's reports: https://sistema.cenipa.fab.mil.br/cenipa/paginas/relatorios/relatorios.php

### 3.2.1 LDA

#### 3.2.1.1 N-gram Analysis

The most frequent n-grams were extracted providing valuable insights into the most common phrases and word combinations in the texts. The analysis of the n-grams obtained was essential to customise the final list of stop words used.

In their study, Kuhn (2018) performed topic modelling to the Aviation Safety Reporting System (ASRS) incident reports and shared the most frequent n-grams found in their dataset. For the purpose of comparison, Table 1 shows the n-grams obtained by Kuhn (2018) and those resulted from this study with the traditional LDA model and their direct translation. The translated version was obtained using the Google Translate[22] resource. Table 1 indicates possible differences on the reporting standardisation. It also shows that removing stop words from the texts directly affects the human-interpretability of the texts in Portuguese. Finally, it indicates the importance given to having a valid Certificate of Airworthiness and Medical Certificate.

**Table 1: n-grams obtained by Kuhn (2018) and LDA model**

| Kuhn (2018) | LDA | English Translation |
|---|---|---|
| | **5-grams** | |
| First officer FO pilot flying | Atividade responsabilidade baseiase combinação objetiva | Activity responsibility based combination objective |
| Cleared visual approach runway R | Baseiase combinação objetiva efeitos adversos | Based combination objective effects adverse |
| Climb via SID except maintain | Responsabilidade baseia-se combinação objetiva efeitos | Responsibility based combination objective effects |
| Declared emergency returned departure airport | Bordo lesões danos ileso leve | Onboard injuries damage unhurt lightweight |
| We cleared visual approach runway | Lesões danos ileso leve grave | Injuries damage unhurt lightweight serious |
| | **4-grams** | |
| First officer pilot flying | Certificado aeronavegabilidade ca válido | Certificate airworthiness ca valid |
| In future I will | Dentro limites peso balanceamento | Inside limits weight balancing |

[2] Google Translate: https://translate.google.com/?sl=en&tl=pt&op=translate

|  |  |  |
|---|---|---|
| I asked first officer | Cadernetas célula motor hélice | Carnets cell engine propeller |
| Cleared visual approach runway | Fogo sobrevivência eou abandono | Fire survival and/or abandonment |
| Aircraft maintenance manual AMM | Certificado medico cma válido | Certificate medical cma valid |
|  | **3-grams** |  |
| Air carrier X | Bordo lesões danos | Onboard injuries damage |
| First officer I | Dentro limites peso | Within weight limits |
| At point I | Certificado aeronavegabilidade ca | Certificate airworthiness ca |
| At time I | Limites peso balanceamento | Limits weight balancing |
| Landed without incident | Aeronavegabilidade ca válido | Airworthiness ca valid |

## 3.3 Data Mining

### 3.3.1 LDA models

A dictionary and a corpus were created from the pre-processed texts. In the context of LDA, the dictionary is the mapping between the words and their IDs, while the corpus is the collection of documents. Each document is represented as a bag-of-words; the order of the words in the document is not relevant (Blei, 2003). The LDA model was performed for different numbers of topics. The optimal models were chosen in regards of the coherence and perplexity scores as it was also the method used by Kim *et al.* (2023). The two metrics were combined in a single value which was called *composite score*. The coherence score measures the connection between the words in a topic. Higher scores indicate higher semantic relationships and, thus, a higher human-interpretability (UCDavis, 2024). On the other hand, the perplexity score measures how well a topic model can predict unseen data. Lower scores indicate that the model is better at predicting the unseen data (UCDavis, 2024). The scores were plotted for better visualisation of the optimal number of topics and their correspondent evaluation metrics.

### 3.3.1.1 Cross-Language Model + LDA

Some visualisation libraries require data to be in a JSON-serialisable format. A function was created to ensure that the output of the model has the correct format for a successful visualisation.

### 3.3.2 Word2Vec + K-means

The method used to perform word2vec and k-means was suggested by Castillo (2018). The tokens generated in the Pre-processing and Transformation phase were utilised to train the word2vec model and create vectors; one vector per document. Each documented was represented as the average of the numerical values of its words.

MiniBatchKMeans was used for the topic modelling approach. The quality of the clusters and optimal number of topics were evaluated using the silhouette coefficient.

### 3.3.3 BERTopic

BERTopic was the third method used for topic modelling. The technique was chosen for being effective with multilingual texts. UMAP is the standard method for dimensionality reduction of the approach (Grootendorst, 2022) and was, therefore, the one applied. The optimal number of topics and evaluation of the model was determined by the coherence score.

## 3.4 Interpretation / Evaluation

### 3.4.1 Latent Dirichlet Allocation (LDA) and variations

Equations 2, 3 and 4 show how the *composite score* was calculated. The metric was used to determine the optimal number of topics. However, the evaluation of the models will be based on the Coherence Score, given that the composite score is not available for BERTopic.

$$normalisedCoherence_i = \frac{C_i - \min(C)}{\max(C) - \min(C)} \tag{2}$$

$$normalisedPerplexity_i = 1 - \frac{P_i - \min(P)}{\max(P) - \min(P)} \tag{3}$$

$$\begin{aligned} &compositeScore_i \\ &\quad = 0{,}5 * normalisedCoherence_i + 0{,}5 * normalisedPerplexity_i \end{aligned} \tag{4}$$

In the visualisation step, the pyLDAvis was used as an interactive solution for the visualisation of the topics.

### 3.4.2 Word2Vec + K-means

The silhouette coefficient in given by Equation 1. Its values vary from -1 to 1. Higher values indicate better cluster separation (Zhao *et al.*, 2018). The parameters a and b are the mean of the distances intra clusters and the “distance between a sample and the nearest cluster that the sample is not a part of”, respectively (scikit-learn developers, 2024).

$$silhouette_i = \frac{b - a}{\max(a, b)} \tag{1}$$

To visualise the word embeddings and document clusters in a 2D-space, t-SNE was employed. Heuer (2015) has shown that, when combined, word2vec and t-SNE allows exploring texts like a geographical map with words with similar meanings being plotted together.

### 3.4.3 BERTopic

The evaluation of the BERTopic model was based on the coherence metric. Additionally, a visual inspection of the topics and their distributions was used as a qualitative evaluation method. In order to analyse the possible existing relationships between topics, a hierarchical clustering dendrogram was plotted.

# 4 Design Specification

This section specifies the architectures and frameworks used in the models implemented. They were designed to meet a series of requirements, i.e., scalability, flexibility, interpretability and reproducibility. The reproducibility is given by setting a seed. The modular architecture guarantees easy substitution or addition of new components. Lastly, the visual tools in place facilitate the interpretation of the results and topics obtained.

The modular architecture is based on five key components: text extraction, text pre-processing, n-gram analysis, topic modelling framework and visualisation layer. The extraction module uses the PyMuPDF (fitz) library to extract the raw text from the PDF files. It is designed to handle eventual errors that could occur during the extraction process. The text pre-processing pipeline employs NLTK resources and regular expressions to process the cleaning and standardisation steps described in the Research Methodology. The n-gram analysis component utilises the NLTK function *ngram* to extract n-grams of different lengths and incorporates a counter in order to determine the frequency of the n-grams obtained.

## 4.1 Latent Dirichlet Allocation (LDA) and variations

The topic modelling framework is based on the implementation of LDA using the Gensim library which allows processing large corpora. The evaluation metrics coherence and perplexity scores were used as parameter tuning to determine the optimal number of topics for the model.

In the visualisation layer, matplotlib and pyLDAvis are used to create plots of the model and performance metrics.

## 4.2 Word2Vec + K-means

The proposed solution employed Gensim word2vec model to generate the word embeddings and MiniBatchKMeans from scikit-learn, aiming an efficient clustering process. The architecture is similar to the one previously described with the topic modelling framework consisting of a word embedding module comprising the model training and evaluation, followed by two other modules: document vectorisation and clustering.

## 4.3 BERTopic

The architecture of the BERTopic model includes a module for dimensionality reduction using UMAP which was configured to reduce dimensions to 2 for visualisation purposes.

UMAP stands for Uniform Manifold Approximation. According to McInnes, Healy and Melville (2020), it has the ability to reduce dimensionality preserving the quality of data.

# 5 Implementation

The implementation of the five models was done in Python and leveraged several libraries for natural language processing. Jupyter Notebook and the integrated development environment (IDE) PyCharm were used for an interactive development and easy visualisation of the results.

The models were implemented following a modular approach. Different functions were defined for the major steps of each pipeline. This design was chosen to allow easy testing and modifications, if necessary. Additionally, tqdm was used where possible to monitor progress.

## 5.1 Latent Dirichlet Allocation (LDA) and variations

The outputs produced by the LDA models were:

1. Pre-processed text data
2. N-gram frequency lists
3. Coherence, perplexity and composite scores for each model
4. Plots of the coherence and perplexity scores
5. Interactive visualisation of the final model with optimal number of topics
6. Saved LDA model file

Figure 2 plots the values of coherence, perplexity and composite scores against their respective number of topics. Figure 3 plots the same described for the LDA model with stemming. Algorithm 1 describes the steps taken to obtain the composite score.

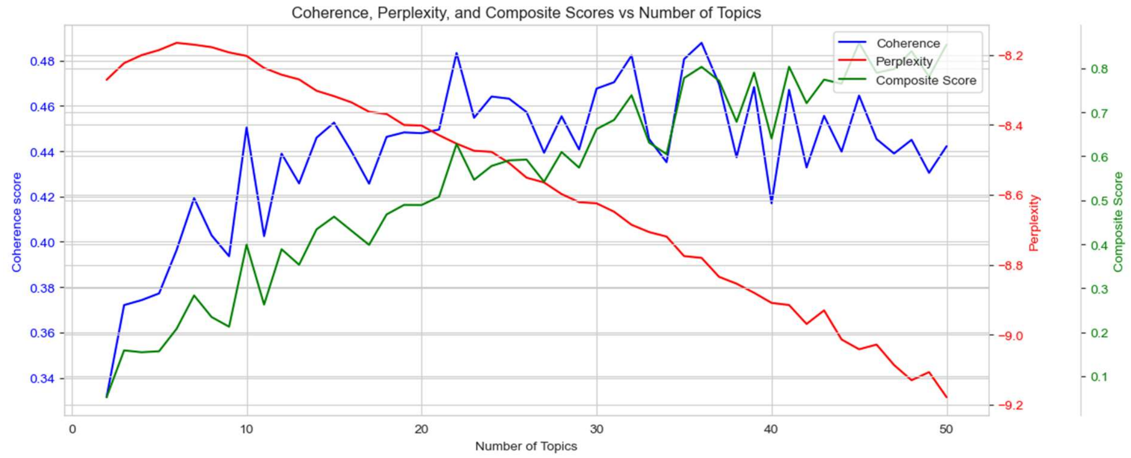

**Figure 2: LDA - Image of Coherence, Perplexity and Composite Scores vs Number of Topics. The best number of clusters is 45 with a correspondent composite score of 0.8580.**

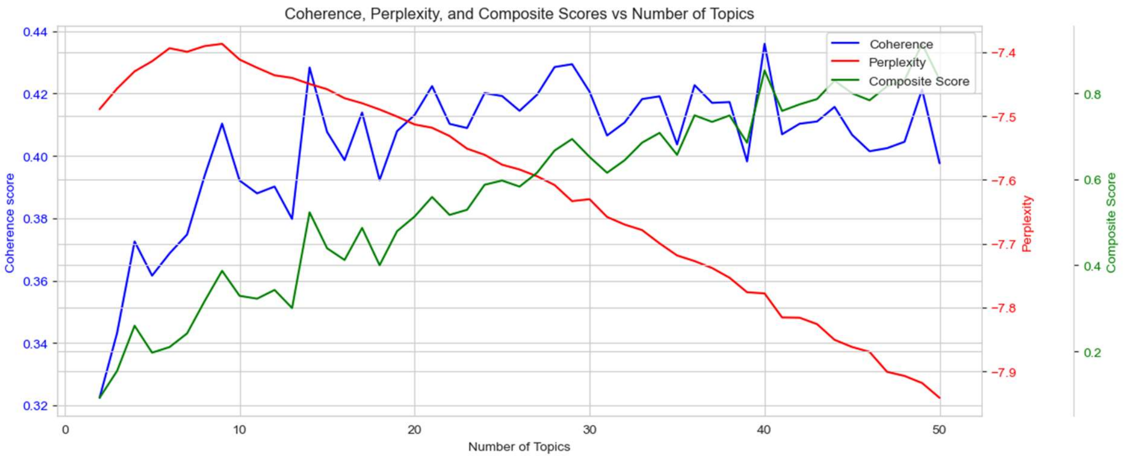

**Figure 3: LDA + Stemming - Image of Coherence, Perplexity and Composite Scores vs Number of Topics. The best number of clusters is 47 with a correspondent composite score of 0.9146.**

**Algorithm 1** Calculating Coherence and Perplexity Scores for different number of topics

```
FUNCTION compute_coherence_and_perplexity(dictionary, corpus, texts, start, limit, step):

        INITIALIZE empty lists for coherence_values, perplexity_values, and model_list

        FOR num_topics FROM start TO limit STEP step:
                CREATE LdaModel with num_topics
                COMPUTE coherence score
                COMPUTE perplexity
                ADD scores and model to respective lists
        RETURN model_list, coherence_values, perplexity_values
```

### 5.1.1 Translation

The translation of the texts was performed leveraging the model *unicamp-dl/translation-pt-en-t5* available on Hugging Face[3]. The model adapted an English tokenizer to handle characters from the Portuguese language. It was chosen for being available for public use and achieving performance compared to state-of-the art-models (Lopes *et al.*, 2020). Figure 4 plots the values of coherence, perplexity and composite scores against their respective number of topics for the cross-language + LDA model.

[3] Hugging Face: https://huggingface.co/unicamp-dl/translation-en-pt-t5

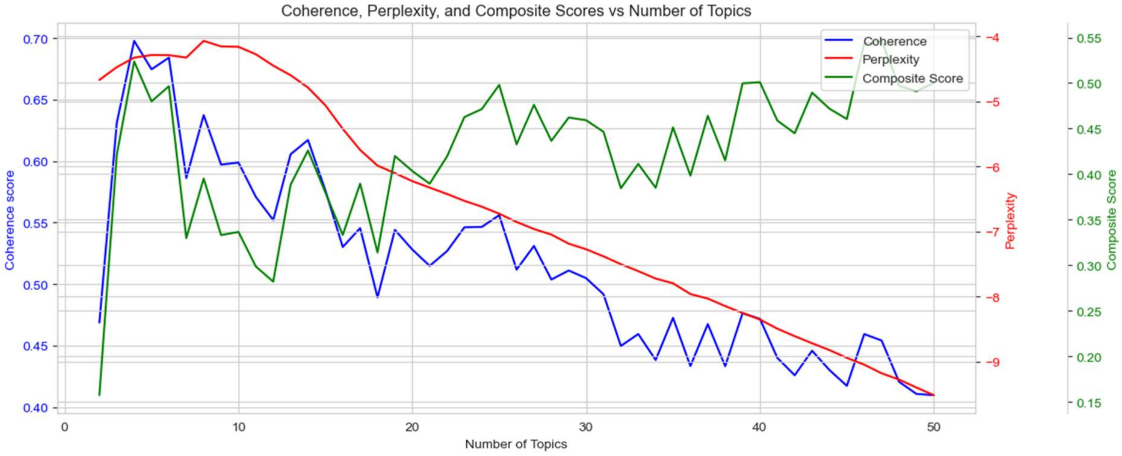

**Figure 4: Cross-language model + LDA - Image of Coherence, Perplexity and Composite Scores vs Number of Topics. The best number of clusters is 45 with a composite score of 0.5463.**

## 5.2 Word2Vec + K-means

Similarly to the LDA model, the wor2vec with k-means has also produced pre-processed text data as an output. The additional key outputs were:

1. Trained word2vec model
2. Document vectors
3. Cluster assignments to topics
4. Visualisations
5. Clusters results
6. Saved word2vec + k-means model

Each document was assigned to a specific topic cluster. The word embeddings and documents clusters were plotted in 2D. For the analysis of the clusters obtained, the top terms of each cluster were determined based on two different parameters: centroids and frequency. Centroids was the approach used by Castillo (2018) who also suggested using the frequency as alternative method. Figure 5 plots the values of the silhouette score for different numbers of clusters.

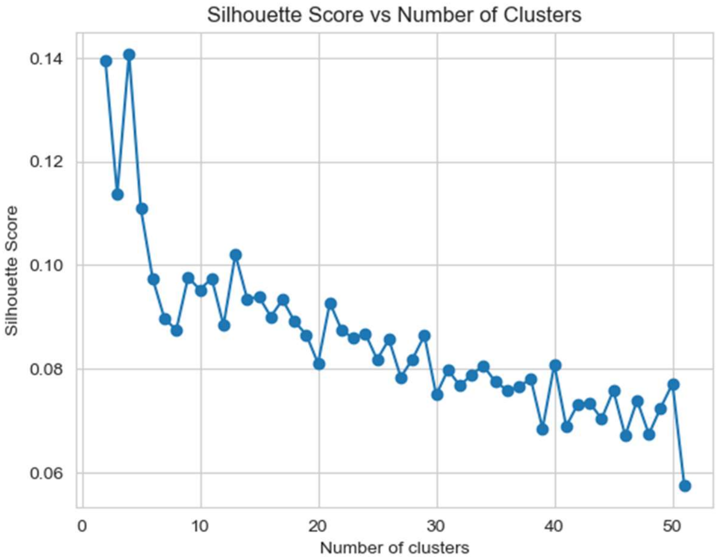

**Figure 5: Word2Vec + K-means - Image of the Silhouette Scores vs Number of Clusters. The best number of clusters is 4 with a correspondent silhouette score of 0.1408.**

## 5.3 BERTopic

The final model with the optimal number of topics was saved using the BERTopic's built-in save method. The outputs produced by the approach were the trained BERTopic model, various visualisations and a data frame containing document embeddings and their assigned topics. Figure 6 shows the coherence score for each corresponding number of topics.

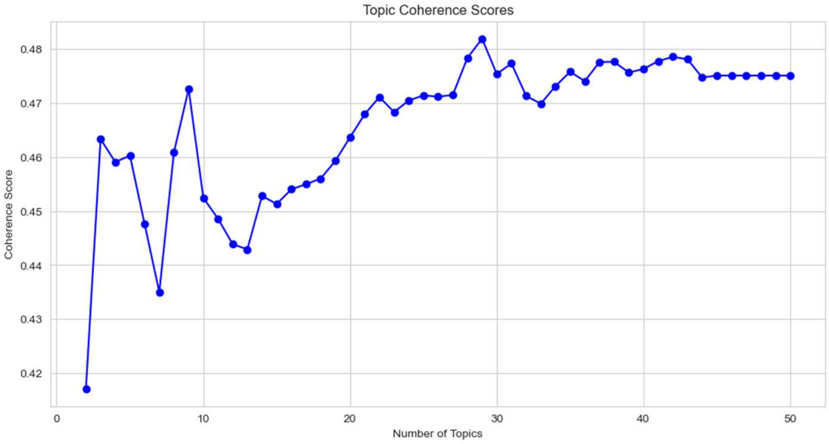

**Figure 6: BERTopic - Image of the Coherence Scores vs Number of Clusters. The best number of topics is 29 with a correspondent coherence score of 0.4819.**

# 6 Evaluation

This section presents the evaluation results for the topic modelling methods explored. The performance of the models was assessed using coherence, perplexity, composite and silhouette scores. Table 2 summarises the performance metrics obtained and the optimal number of topics for each of the methods tested.

**Table 2: Models and their respective evaluation scores**

| | Number of Topics | Coherence Score | Perplexity Score | Composite Score | Silhouette Score |
|---|---|---|---|---|---|
| LDA | 45 | 0.4645 | 9.0417 | 0.8580 | - |
| LDA + Stemming | 49 | 0.4213 | 7.9184 | 0.9146 | - |
| Translation + LDA | 47 | 0.4542 | 9.1778 | 0.5463 | - |
| Word2Vec + K-means | 4 | - | - | | 0.1408 |
| BERTopic | 29 | 0.4819 | - | | - |

The traditional LDA approach showed a good performance. It identified 45 topics with a coherence score of 0.4645. The composite score of 0.8580 suggests a good balance between coherence and perplexity. Applying stemming to LDA resulted in an optimal number of topics of 49. This method achieved the best composite score among the LDA variations, 0.9146. Even though the model presents good generalisation capabilities, its coherence score was the lowest observed, demonstrating reduced topic-interpretability. The cross-language model identified 47 as optimal number of topics with a coherence score of 0.4542. This model was the LDA variation with lowest performance, achieving 0.5463 as composite score. Language-specific information might have been lost with the translation process.

Although the comparison between the silhouette score and coherence score is not straightforward as the metrics use different scales and measure different aspects of the models, it is clear the disparity between the number of optimal topics identified. The approach resulted in 4 topics, with a silhouette score of 0.1408 which indicates a poor cohesion and separation among the clusters. This can also be seen in Figure 7 that illustrates the 2D visualisation of the clusters obtained.

BERTopic, the last method performed, identified 29 topics. The technique achieved the highest coherence score, 0.4819, indicating higher human-interpretability. The model outperformed all the LDA approaches in terms of coherence.

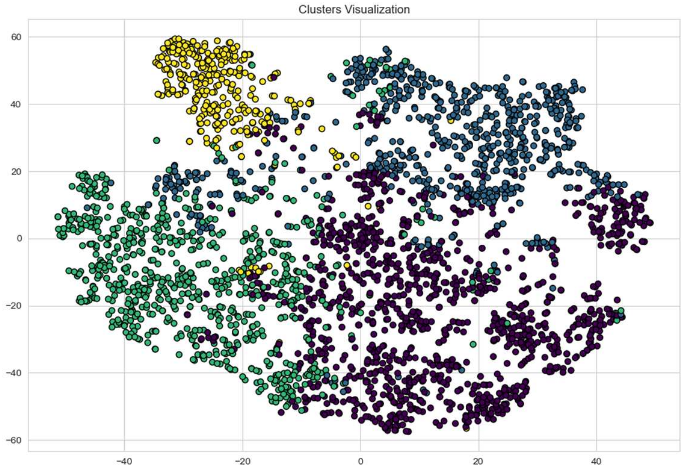

**Figure 7: Word2Vec + K-means - The four clusters found by the word2vec + k-means model is represented in the image. It is notable that the clusters differ in size.**

## 6.1 Discussion

Based on the composite scores, the LDA with stemming approach shows the best performance among the LDA variations. It also presents the best perplexity score. BERTopic is a superior approach in regards of coherence score, which suggests that its topics are more interpretable.

The traditional LDA method shows a good balance between coherence and perplexity, ranking second if considering the composite score as the parameter for comparison. The cross-language + LDA approach maintained a good coherence score, but shows the lowest composite score among the LDA models.

The variation among the number of topics identified by the different techniques explored and the existing topics, suggests that further qualitative analysis could contribute to the identification of the best granularity to classify the occurrences reported. An interesting observation is that although BERTopic has identified 29 themes within the texts analysed, its dendrogram illustrated in Figure 8 shows four main groups of topics, same number obtained by the word2vec approach with k-means. The method might not be the best option to classify the documents, but it could be used in cases where texts needed to be clustered in more generic groups.

The lack of metrics to compare across all models limit the evaluations that can be made. In addition, qualitative analysis could contribute to deeper insights on the quality of the topics. Another aspect worth noting is that comparing the performance of the models for the same number of topics was not explored and this could influence the results of performance obtained.

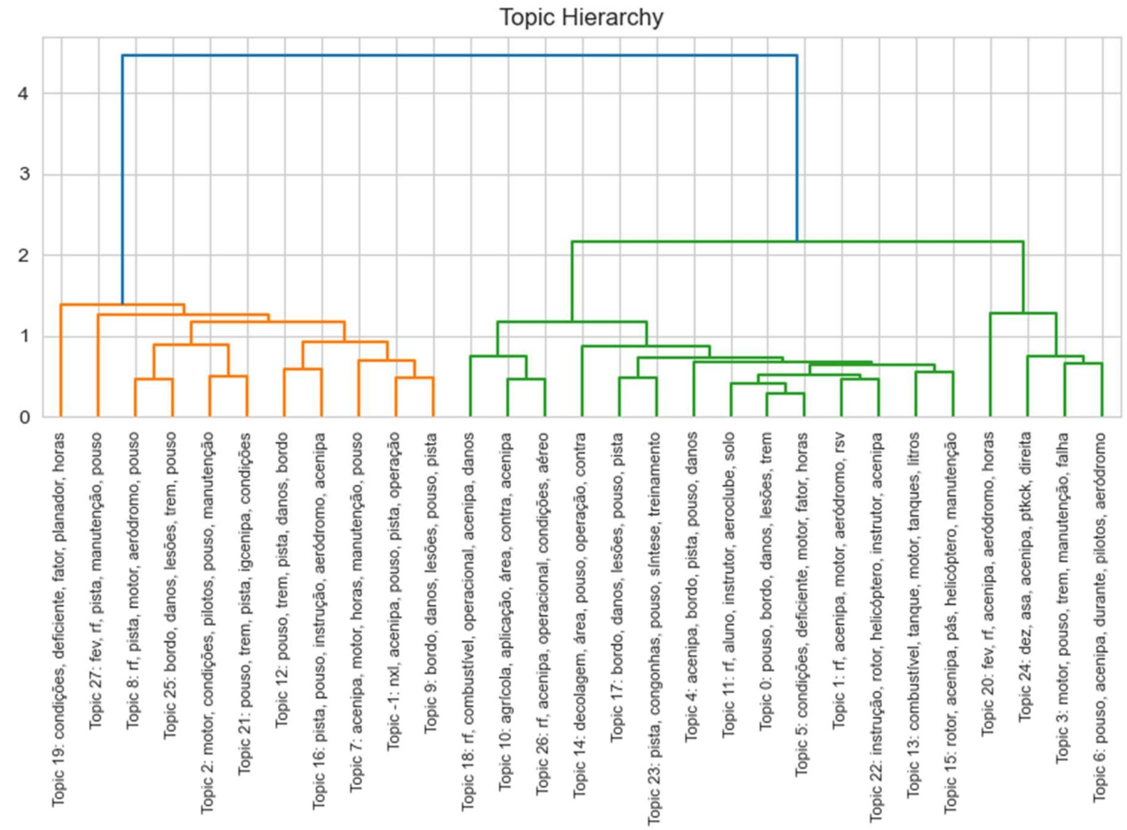

**Figure 8: BERTopic - Image of dendrogram showing topics hierarchy**

# 7 Conclusion and Future Work

This study aimed to evaluate and compare how effective different topic modelling approaches are when applied to aviation safety reports in Brazilian Portuguese. The dataset utilised has not been used in previous research, and this gap is now addressed. Five different models were implemented, namely Latent Dirichlet Allocation (LDA), LDA with stemming, a cross-language model which used translation followed by LDA, word2vec combined with k-means and BERTopic. Each of the models was evaluated by an appropriate score.

BERTopic has shown the best performance based on the coherence score, achieving 0.4819. Among the LDA variations, LDA with stemming obtained the highest composite score, a metric calculated using both coherence and perplexity scores. The approach using word2vec with k-means appears to be more suitable for categorisations where a detailed topic analysis is not necessary. Additionally, the low composite score obtained by the cross-language model indicates a potential information loss consequence of the translation process.

Choosing the appropriate technique when performing topic modelling is essential to obtain satisfactory results. Characteristics of the data and the granularity desired must be taken into consideration during the decision-making process. Furthermore, this study indicates potential benefits of using more advanced techniques such as BERTopic to analyse aviation safety data.

Future work could build upon the results obtained and explore a qualitative analysis of the topics identified by the different models. It could apply different tools available for the Portuguese language. In addition, a more ambitious study could work on developing a domain-specific pre-trained model for the Portuguese language.

# References

Alammar, J. and Grootendorst, M. (2024) *Hands-On Large Language Models*. O'Reilly Media, Inc. Available at: https://learning.oreilly.com/library/view/hands-on-large-language/9781098150952/ (Accessed: 8 July 2024).

Baishya, D. and Baruah, R. (2022) 'Recent Trends in Deep Learning for Natural Language Processing and Scope for Asian Languages', in *2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*. *2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, Trichy, India: IEEE, pp. 408–411. Available at: https://doi.org/10.1109/ICAISS55157.2022.10010807.

Blei, D.M. (no date) 'Latent Dirichlet Allocation'. Available at: https://jmlr.org/papers/volume3/blei03a/blei03a.pdf (Accessed: 8 July 2024).

Buselli, I. *et al.* (2022) 'Natural language processing for aviation safety: extracting knowledge from publicly-available loss of separation reports,' *Open Research Europe*, 1, p. 110. https://doi.org/10.12688/openreseurope.14040.2.

Castillo, D. (2018) *How to Cluster Documents Using Word2Vec and K-Means*. Available at: https://dylancastillo.co/posts/nlp-snippets-cluster-documents-using-word2vec.html (Accessed: 8 July 2024).

Churchill, R. and Singh, L. (2021) 'The Evolution of Topic Modeling', *ACM Computing Surveys*, 54(10s), pp. 1–35. Available at: https://doi.org/10.1145/3507900.

De Miranda, G.R., Pasti, R. and De Castro, L.N. (2020) 'Detecting Topics in Documents by Clustering Word Vectors', in F. Herrera, K. Matsui, and S. Rodríguez-González (eds) *Distributed Computing and Artificial Intelligence, 16th International Conference*. Cham: Springer International Publishing (Advances in Intelligent Systems and Computing), pp. 235–243. Available at: https://doi.org/10.1007/978-3-030-23887-2_27.

Fayyad, U.M., Piatetsky-Shapiro, G. and Smyth, P. (1996) 'From data mining to knowledge discovery in databases,' *AI Magazine*, 17(3), pp. 37–54. https://doi.org/10.1609/aimag.v17i3.1230.

Ferreira, J., Gonçalo Oliveira, H. and Rodrigues, R. (2019) 'Improving NLTK for Processing Portuguese', *OASIcs, Volume 74, SLATE 2019*, 74, p. 18:1-18:9. Available at: https://doi.org/10.4230/OASICS.SLATE.2019.18.

Grootendorst, M. (2022) 'BERTopic: Neural topic modeling with a class-based TF-IDF procedure'. arXiv. Available at: http://arxiv.org/abs/2203.05794 (Accessed: 2 August 2024).

Haider, M.M. *et al.* (2020) ‘Automatic Text Summarization Using Gensim Word2Vec and K-Means Clustering Algorithm’, in *2020 IEEE Region 10 Symposium (TENSYMP). 2020 IEEE Region 10 Symposium (TENSYMP)*, Dhaka, Bangladesh: IEEE, pp. 283–286. Available at: https://doi.org/10.1109/TENSYMP50017.2020.9230670.

Heuer, H. (2016) 'Text comparison using word vector representations and dimensionality reduction,' *arXiv (Cornell University)* [Preprint]. https://doi.org/10.48550/arxiv.1607.00534.

Kim, J.H. *et al.* (2023) ‘Aviation Safety Mandatory Report Topic Prediction Model using Latent Dirichlet Allocation (LDA)’, *Journal of the Korean Society for Aviation and Aeronautics*, 31(3), pp. 42–49. Available at: https://doi.org/10.12985/ksaa.2023.31.3.042.

Kuhn, K.D. (2018) ‘Using structural topic modeling to identify latent topics and trends in aviation incident reports’, *Transportation Research Part C: Emerging Technologies*, 87, pp. 105–122. Available at: https://doi.org/10.1016/j.trc.2017.12.018.

Lopes, A. *et al.* (2020) 'Lite training Strategies for Portuguese-English and English-Portuguese translation,' *arXiv (Cornell University)* [Preprint]. https://doi.org/10.48550/arxiv.2008.08769.

Luo, Y. and Shi, H. (2019) ‘Using lda2vec Topic Modeling to Identify Latent Topics in Aviation Safety Reports’, in *2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS). 2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS)*, Beijing, China: IEEE, pp. 518–523. Available at: https://doi.org/10.1109/ICIS46139.2019.8940271.

Maschek, A. and Stöckl, A. (2023) ‘Automated evaluation of sport and leisure accident reports using natural language processing’, in *2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME). 2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, Tenerife, Canary Islands, Spain: IEEE, pp. 1–5. Available at: https://doi.org/10.1109/ICECCME57830.2023.10252673.

McInnes, L., Healy, J. and Melville, J. (2020) ‘UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction’. arXiv. Available at: http://arxiv.org/abs/1802.03426 (Accessed: 11 August 2024).

Rose, R.L. *et al.* (2022) ‘Application of structural topic modeling to aviation safety data’, *Reliability Engineering & System Safety*, 224, p. 108522. Available at: https://doi.org/10.1016/j.ress.2022.108522.

scikit-learn developers (2024) *silhouette_score*, *scikit learn*. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html (Accessed: 8 November 2024).

Tikayat Ray, A. *et al.* (2023) ‘Examining the Potential of Generative Language Models for Aviation Safety Analysis: Case Study and Insights Using the Aviation Safety Reporting System (ASRS)’, *Aerospace*, 10(9), p. 770. Available at: https://doi.org/10.3390/aerospace10090770.

UCDavis (2024) *Topic Modeling*, *UCDavis DataLab*. Available at: https://ucdavisdatalab.github.io/workshop_nlp_reader/chapters/06_topic-modeling.html (Accessed: 8 October 2024).

Wang, W. (2023) ‘Different natural languages, equal importance’, *Patterns*, 4(8), p. 100821. Available at: https://doi.org/10.1016/j.patter.2023.100821.

Xing, Y. *et al.* (2024) ‘Discovering latent themes in aviation safety reports using text mining and network analytics’, *International Journal of Transportation Science and Technology*, p. S2046043024000297. Available at: https://doi.org/10.1016/j.ijtst.2024.02.009.

Zhao, S. *et al.* (2018) ‘Silhouette Scores for Arbitrary Defined Groups in Gene Expression Data and Insights into Differential Expression Results’, *Biological Procedures Online*, 20(1), p. 5. Available at: https://doi.org/10.1186/s12575-018-0067-8.