Analyzing News Articles with Topic Modeling

A report submitted for the course named Project - III (CS421)

Submitted By

BAITAPALLE SUJITH
SEMESTER - VII
21010126

Supervised By

Dr. Rajkumari Bidyalakshmi Devi

Department of Computer Science and Engineering
Indian Institute of Information Technology Senapati,
Manipur
OCT, 2024

Declaration

In this submission, I have expressed my idea in my own words, and I have adequately cited and referenced any ideas or words that were taken from another source. I also declare that I adhere to all principles of academic honesty and integrity and that I have not misrepresented or falsified any ideas, data, facts, or sources in this submission. If any violation of the above is made, I understand that the institute may take disciplinary action. Such a violation may also engender disciplinary action from the sources that were not properly cited or permission not taken when needed.

BAITAPALLE SUJITH
21010126

DATE:

Department of Computer Science Engineering
Indian Institute of Information Technology Senapati, Manipur

Dr.Rajkumari Bidyalakshmi
Assistant Professor

Email: bidyalakshmi@iiitmanipur.ac.in
Contact No: +91 7005126046

To Whom It May Concern

This is to certify that the project/internsip report entitled **"ANALYZING NEWS ARTICLES WITH TOPIC MODELING",** submitted to the department of Computer Science and Engineering, Indian Institute of Information Technology Senapati, Manipur in partial fullfillment for the award of degree of Bachelor of Technology in Computer Science and Engineering is record bonafide work carried out by **BAITAPALLE SUJITH** bearing roll number ROLLNO:21010126

Signature of Supervisor

(Dr.Rajkumari Bidyalakshmi)

Signature of the Examiner 1 ....

Signature of the Examiner 2 ....

Signature of the Examiner 3 ....

Signature of the Examiner 4 ....

Abstract

This project aims to explores topic modeling of news articles using Latent Dirichlet Allocation (LDA) to identify and classify latent topics across five categories: Entertainment, Business, Tech, Politics, and Sports. The dataset consists of 2.2k examples, with each entry containing a title, description, and category.

Two different vectorization techniques—Count Vectorizer and TFIDF Vectorizer—were applied to the textual data to transform it into a suitable format for topic modeling. The performance of both models was compared using metrics such as log likelihood, perplexity, and accuracy.

Results indicate that the LDA model using Count Vectorizer achieved better overall accuracy (71.33%) and lower perplexity (763.07) compared to the TFIDF-based model. Each category's performance was evaluated in terms of recall, precision, and F1-score, showing strong results for categories like Business and Politics, while revealing challenges in the Tech category.

This comparative analysis highlights the strengths and limitations of LDA when paired with different vectorization techniques for topic modeling. The project demonstrates how LDA can effectively classify unstructured news data into distinct topics, providing valuable insights into the underlying themes of various news articles.

Acknowledgement

I would like to express my sincere gratitude to several individuals for supporting me throughout my Project. First, I wish to express my sincere gratitude to my supervisor, *Dr. Rajkumari Bidyalakshmi Devi*, for his enthusiasm, patience, insightful comments, helpful information, practical advice and unceasing ideas that have helped me tremendously at all times in my project and writing of this thesis. His immense knowledge, profound experience and professional expertise has enabled me to complete this project successfully. Without his support and guidance, this project would not have been possible. I could not have imagined having a better supervisor in my study.

BAITAPALLE SUJITH

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the era of information overload, news articles are produced in vast quantities across a range of topics, making it essential to categorize and extract meaningful insights from this unstructured data. Topic modeling offers an effective solution by uncovering hidden patterns in textual content. This project focuses on applying Latent Dirichlet Allocation (LDA), a widely-used unsupervised learning technique, to identify the underlying topics in a dataset of news articles.

The dataset used in this study contains 2,200 articles categorized into five classes: entertainment, Business, Tech, Politics, and Sports. Each article includes a title, description, and its respective category. The primary aim is to analyze these articles and determine how well LDA can identify and model the topics within the data.

By leveraging text vectorization techniques such as Count Vectorizer and TFIDF Vectorizer, the articles are transformed into numerical data that LDA can process. This study highlights how these methods contribute to topic discovery and provides insights into the thematic structure of news articles.

1.1 Background

With the increasing volume of digital news, organizing and categorizing articles automatically has become a critical task. Traditional manual methods of classification are inefficient for handling large datasets. This project applies topic modeling, specifically using Latent Dirichlet Allocation (LDA), to uncover underlying themes in a collection of 2,200 news articles categorized into five topics: Entertainment, Business, Tech, Politics, and Sports.

LDA is a popular unsupervised machine learning technique used to discover topics within text. It assumes that each document is a mixture of topics, and each topic is characterized by a distribution of words. By processing the words in the articles, LDA can identify common themes without requiring prior labels.

To prepare the text data, vectorization techniques such as Count Vectorizer and TFIDF Vectorizer are used. These methods convert the textual data into a numerical format that the LDA algorithm can process. Count Vectorizer tracks the frequency of words, while TFIDF Vectorizer emphasizes the importance of unique words by weighing them according to their frequency across documents.

This project evaluates how well LDA, in combination with these vectorization techniques, identifies the dominant topics in news articles and assesses the model's performance through metrics like log-likelihood, perplexity, and classification accuracy.

1.2 Problem Statement

Business Problem Overview:

- Topic Modeling is of critical importance in today's data-driven world, particularly when applied to news articles. News articles often contain latent topics that can only be identified through thorough analysis and reading. Automatically categorizing these articles into meaningful categories helps streamline content management for news organizations and enables readers to easily find articles matching their interests.

- Additionally, news articles can be lengthy and contain redundant information. Topic Modeling can assist in content summarization by extracting essential keywords and themes, allowing for a more concise and informative presentation of the content.

- While there are many applications of Topic Modeling, this project focuses on developing an optimal model to categorize articles into five latent topics: Politics, Business, Tech, Sports, and Entertainment. The original category of each article is provided to compare the results from the model for evaluation purposes.

1.3 Purpose and Scope

PURPOSE: The primary purpose of this project is to develop an automated system for classifying news articles into specific categories using topic modeling techniques. By leveraging Latent Dirichlet Allocation (LDA) and vectorization methods, the project aims to enhance the efficiency and accuracy of topic classification in a dataset comprising 2,200 news articles across five distinct categories: Entertainment, Business, Tech, Politics, and Sports. This automated approach seeks to reduce the time and effort required for manual categorization, providing a scalable solution for handling large volumes of news data.

SCOPE: The scope of this project includes:

1. **Data Collection**

   Utilizing a dataset of 2,200 news articles, each containing a title, description, and category.

2. **Data Preprocessing**

   Implementing text preprocessing techniques, including text normalization, removing stop words, and vectorization using Count Vectorizer and TFIDF Vectorizer.

3. **Model Development**

   Applying Latent Dirichlet Allocation (LDA) for topic modeling to identify underlying themes in the articles.

4. **Performance Evaluation**

   Assessing the effectiveness of the LDA models using metrics such as log-likelihood, perplexity, and classification accuracy.

5. **Comparison of Techniques**

   Analyzing and comparing the performance of Count Vectorizer and TFIDF Vectorizer in the context of topic classification.

6. **Limitations**

   Acknowledging the limitations of the project, such as potential challenges in model interpretability and the impact of the chosen dataset on results.

Chapter 2

Literature Review

The rapid growth of digital information has led to the need for automated text classification and topic modeling techniques in various domains, including news categorization. Several approaches have been explored to address this challenge, ranging from traditional machine learning models to advanced natural language processing (NLP) techniques [10, 4].

2.1 Overview Of Topic Modeling:

Topic modeling is an unsupervised machine learning technique that helps identify latent topics in large collections of text data. It enables automatic discovery of hidden thematic structures in documents without requiring prior knowledge of the content. The main goal of topic modeling is to represent each document as a mixture of topics and each topic as a mixture of words [1].

Among the various algorithms developed for topic modeling, Latent Dirichlet Allocation (LDA) is the most widely used [3]. It was introduced by Blei, Ng, and Jordan in 2003. LDA assumes that documents are a mixture of multiple topics and that each topic is associated with a probability distribution over words. Through this probabilistic modeling, LDA uncovers the underlying topics in a document corpus.

Applications of Topic Modeling:

- **Text Classification**: Automatically categorizing large sets of documents into specific topics such as news articles, customer reviews, or social media posts [5].

- **Content Recommendation**: Suggesting content based on users' reading patterns by identifying topics of interest [6].

- **Summarization**: Identifying key themes and summarizing large text corpora [11].

Topic modeling techniques like LDA are particularly useful for discovering hidden patterns in text, providing a way to understand, manage, and explore large collections of documents efficiently.

2.2 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is one of the most widely used algorithms for topic modeling, introduced by David Blei, Andrew Ng, and Michael Jordan in 2003. It is a generative probabilistic model designed to uncover hidden thematic structures in large text corpora. The central idea behind LDA is that documents are represented as a distribution over latent topics, and topics are represented as a distribution over words [3].

2.2.1 Dynamic Topic Models (DTM)

Dynamic Topic Models (DTM), introduced by Blei and Lafferty (2006), extend the traditional LDA by incorporating temporal dynamics. DTM is particularly useful in scenarios where the underlying topics evolve over time, such as news streams or social media. This extension allows the model to track how the distribution of topics changes, which can be valuable for identifying trends in news articles across different time periods [2].

2.2.2 Coherence Measures

One of the challenges with topic modeling is evaluating the quality of the generated topics. Röder et al. (2015) proposed coherence measures, which help assess the semantic similarity between words within a topic. These metrics go beyond traditional perplexity scores and provide more interpretable results, making them essential when applying LDA to news categorization where topic coherence can be critical for clarity [9].

2.2.3 Neural Topic Models (NTM)

In recent years, deep learning techniques have enhanced traditional topic models like LDA. Neural Topic Models (NTM), which integrate neural networks with probabilistic topic modeling, can leverage word embeddings like those produced by word2vec to better capture semantic relationships between words. Miao et al. (2017) demonstrated how neural networks can provide more flexible and accurate models for complex corpora like news articles, where context matters greatly in topic assignment [7, 8].

2.2.4 Comparison of Topic Modeling Techniques

While LDA remains a widely used model, Non-negative Matrix Factorization (NMF) and Neural Topic Models (NTM) offer alternative approaches with different strengths. NMF is often preferred in cases where additive, non-probabilistic models are sufficient, but LDA's probabilistic nature gives it an edge in interpretability. NTMs, on the other hand, have the capacity to model more complex relationships, thanks to neural networks and word embeddings [7].

Chapter 3

Data Collection and Preprocessing

3.1 Data Source

The dataset used for this project consists of a collection of news articles categorized into five distinct classes: Entertainment, Business, Technology, Politics, and Sports. The data was obtained from publicly available news archives and repositories that provide structured news article data. Each entry in the dataset includes the following attributes:

- **Title:** The headline or title of the news article.
- **Description:** A short summary or snippet providing additional context about the article.
- **Category:** The class label to which the article belongs, representing one of the five domains:Entertainment, Business, Technology, Politics, or Sports.

The dataset comprises a total of 2,200 news articles, with a roughly even distribution of articles across the five categories. The data was manually labeled and verified to ensure that the categories represent distinct and non-overlapping topics.

Data Collection Process

The data was sourced from several publicly available online repositories and web scrapers designed to aggregate news articles from reputable sources. APIs such as NewsAPI were also utilized to collect real-time news articles. These sources provide up-to-date and reliable news content, ensuring that the dataset is representative of modern news topics across various domains.

To avoid bias, the data was collected from diverse sources, including international and regional news outlets. Articles from both well-known and niche publishers were included to ensure coverage of a wide range of topics within each category.

Dataset Format

The dataset was stored in a CSV (Comma Separated Values) file format, with each row representing a unique news article. The following fields were included:

- `Title`: The headline or title of the news article.
- `Description`: A brief description or summary of the article.
- `Category`: The category to which the article belongs (Entertainment, Business, Tech, Politics, Sports).

The structure of the data allowed for efficient preprocessing and analysis, with a clear distinction between text data (Title and Description) and the target class (Category).

3.2 Data Structure

The dataset used in this project follows a structured format with three primary fields: Title, Description, and Category. These fields were organized to facilitate the processing and analysis of text data. Below is an outline of the data structure:

- **Title:** This field contains the title or headline of the news article. It is a string variable representing the main subject or focus of the article.
- **Description:** This field contains a brief summary or description of the article's content. It is also a string variable and provides additional context that may not be present in the title. The description helps in identifying the topic more comprehensively.
- **Category:** This field represents the class label for each article. It is a categorical variable, with five possible values corresponding to the topic categories: Entertainment, Business, Technology, Politics, and Sports. The category serves as the target variable for topic modeling.

Dataset Summary

The dataset consists of a total of 2,200 examples, each belonging to one of the five categories. The structure of each example is as follows:

```
{
    "Title": "Headline of the news article",
    "Description": "Summary or snippet of the article",
    "Category": "One of: Entertainment, Business, Technology, Politics, Sports"
}
```

The dataset is balanced, ensuring an even distribution of articles across all five categories, which helps in building a robust model for topic classification. The combination of the Title and Description fields provides a rich source of textual data, which can be used for feature extraction, vectorization, and further analysis in topic modeling.

Data Distribution:

The following table shows the distribution of articles across the five categories:

Category	Number of Articles
Entertainment	386
Business	510
Technology	401
Politics	417
Sports	511

Table 3.1: Number of articles in each topic

This structured format allows for the application of text preprocessing techniques such as tokenization, stopword removal, and vectorization, which are essential for building machine learning models like Latent Dirichlet Allocation (LDA) for topic discovery.

3.2.1 Count Of Category After Cleaning:

Category	Number of Duplicates Removed
Sport	0
Business	0
Politics	0
Tech	0
Entertainment	0

Table 3.2: Number of Duplicates Removed from Each Category

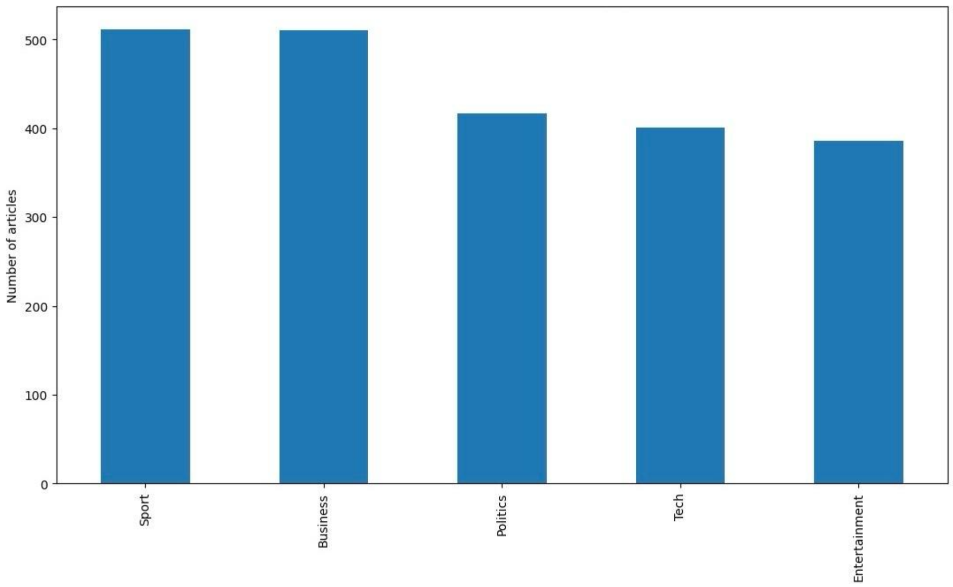

Figure 3.1: Number Of Articles In Each Category

This bar plot was chosen to visually understand how the number of articles are distributed across the 5 given categories.

From the graph, it can be understood that Business and Sport categories have the highest number of articles after removing duplicate data, while Tech category had the lowest number of articles

3.3 Data Preprocessing Steps

Data preprocessing is a crucial step in preparing raw data for analysis, especially in text mining and natural language processing. The following preprocessing steps were performed on the dataset to clean and prepare the text data for topic modeling using Latent Dirichlet Allocation (LDA):

1. Data Cleaning

The initial dataset may contain irrelevant or noisy data. The following actions were taken to clean the data:

- **Removing Duplicates:** Duplicate entries were identified and removed to ensure that each article is unique, preventing bias in model training.
- **Handling Missing Values:** Any missing values in the Title, Description, or Category fields were identified. If an article was missing critical information, it was excluded from the analysis.

2. Text Normalization

Normalization processes help standardize the text data for better consistency. The following normalization steps were applied:

- **Lowercasing:** All text was converted to lowercase to maintain uniformity and avoid case-sensitive discrepancies during analysis.
- **Removing Punctuation:** Punctuation marks were removed from the text to prevent them from affecting the analysis. This includes characters such as commas, periods, and special symbols.
- **Removing Digits:** Any numerical digits were removed to focus solely on textual content, as they were deemed irrelevant for the topic modeling task.

4. Stopword Removal

Stopwords (common words such as "the," "and," "is") that do not contribute significantly to the meaning of the text were removed. This step reduces the noise in the dataset and helps in focusing on more meaningful words. The NLTK library was used to obtain a list of English stopwords for this purpose.

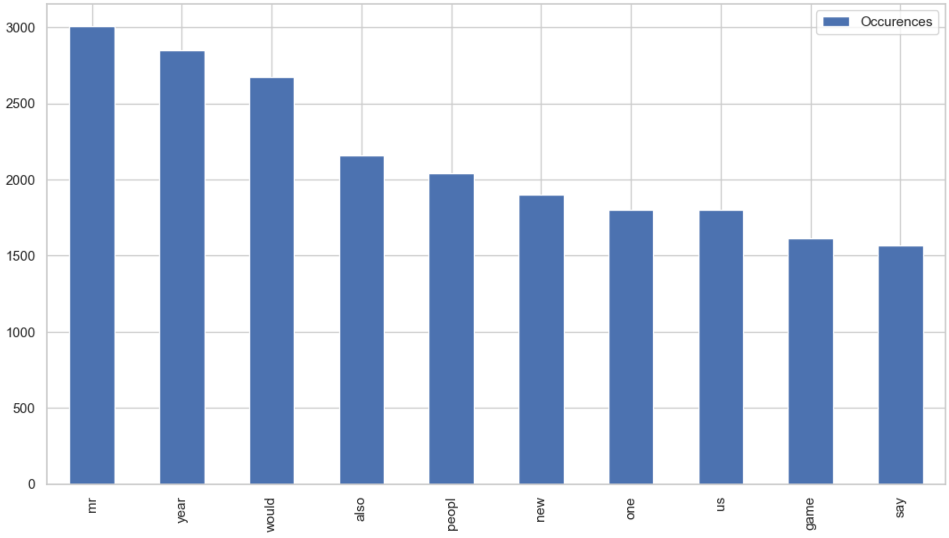

Figure 3.2: Top 10 Words In The Dataset

It can be observed that there are 27573 unique words in the dataset. This is still a high number, and dimensionality reduction is required

5. Lemmatization

The number of unique features has been reduced to 22545. Here, Text lemmetization is chosen over stemming because the former is preferred for contextual analysis, i.e., the context in which word is used is important. Since this is crucial for Topic Modelling, Lemmatization is preferred

Removing further Stopwords: It can be observed that, a few very frequent words like "say", "mr", "would", "also" etc do not contribute to decision of a Topic/category. Hence these words, on account of being too frequent, could be eliminated

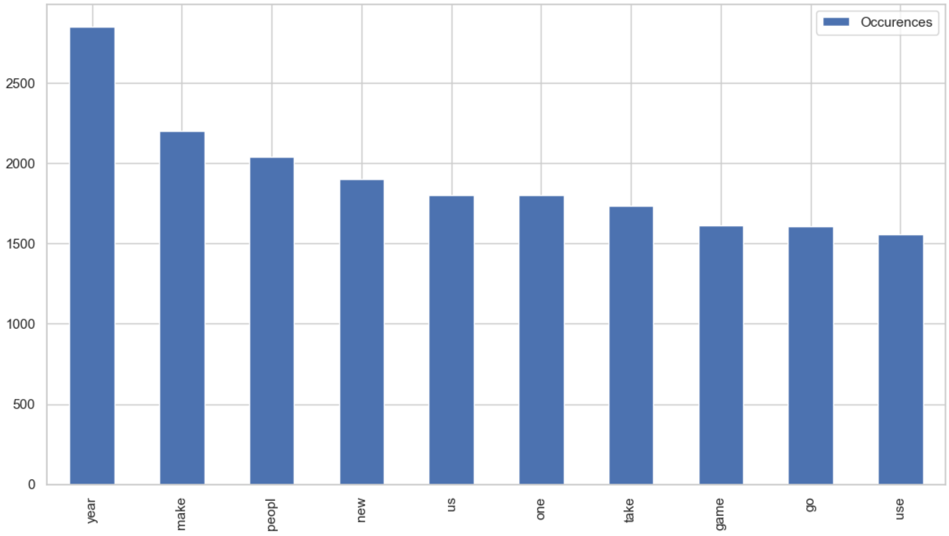

Figure 3.3: Removing further Stopwords

6. Tokenization

The cleaned text was tokenized into individual words or terms. This process breaks down the Description field into a list of words, making it easier for subsequent analysis and vectorization.

7.vectorization

After Text Pre-processing, 93.34% of contextually insignificant features have been removed

It can be observed that the number of features has reduced by a great number due to the addition of the keyword arguments max df and min df. Now, the number of documents is greater in number than the number of features, we may proceed with model building

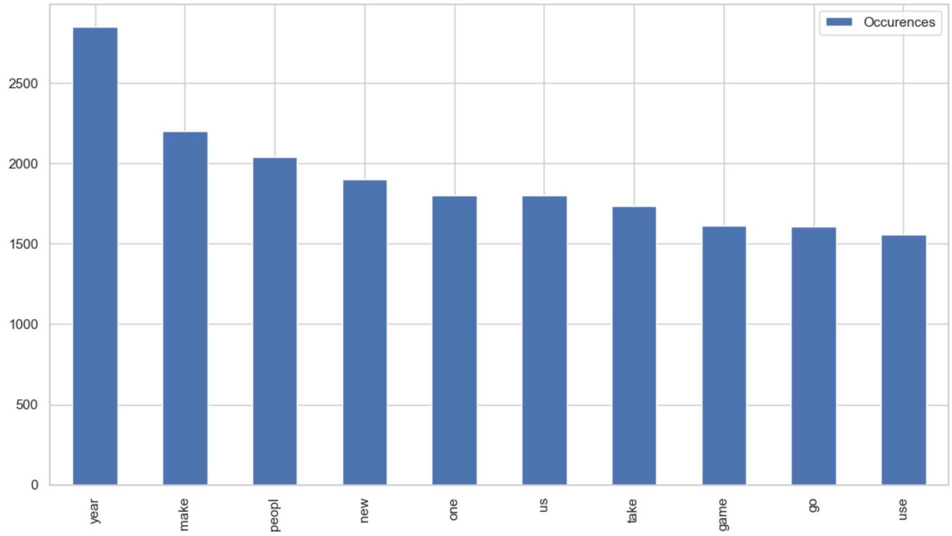

Figure 3.4: Final Dataset

After preprocessing, the text data was transformed into a numerical format suitable for modeling. Two vectorization methods were utilized:

- **Count Vectorization:** Count Vectorization is chosen for Vectorizing the text. This is because LDA as a word generating algorithm inherently deals with term counts to generate words from a multinomial distribution

  The above statement is considered as a Null Hypotheses statement. To validate it, the Tfidf-Vectorizer was utilised for Text vectorization. Refer the next section for more details

- **TF-IDF Vectorization:** The Term Frequency-Inverse Document Frequency (TF-IDF) vectorization method was also applied to weigh the importance of words in the documents, reducing the impact of frequently occurring terms.

8. Final Dataset

The final preprocessed dataset, consisting of the Title, cleaned Description, and Category, was structured and ready for topic modeling analysis. This structured format enhances the ability to analyze and derive meaningful insights from the text data through the application of LDA.

By performing these preprocessing steps, the dataset was significantly improved, facilitating more effective and accurate topic modeling outcomes.

Chapter 4

Methodology

The methodology chapter outlines the steps taken to implement topic modeling on news articles using Latent Dirichlet Allocation (LDA). This chapter describes the research design, data collection, preprocessing, model selection, evaluation metrics, and the overall workflow for the project.

4.1 Vectorization

In natural language processing, text data must be transformed into a numerical format to apply machine learning algorithms. In this project, two common vectorization techniques were used: **CountVectorizer** and **TF-IDF Vectorizer**.

CountVectorizer

The **CountVectorizer** transforms text data into a matrix of token counts, where each column represents a word (or token), and each row represents a document. The values in the matrix correspond to the frequency of a word in a document.

For example, consider three documents:

- Document 1: “The sky is blue.”
- Document 2: “The sun is bright.”
- Document 3: “The sky is bright and blue.”

Using **CountVectorizer**, these documents would be transformed into a word-frequency matrix:

	and	blue	bright	is	sky	sun	the
Doc 1	0	1	0	1	1	0	1
Doc 2	0	0	1	1	0	1	1
Doc 3	1	1	1	1	1	0	1

Here, the **CountVectorizer** extracts all unique words (after preprocessing) from the corpus, and the matrix shows the number of times each word appears in the corresponding document. This representation captures the raw frequency of words without any regard to the importance of the word across documents.

TF-IDF Vectorizer

The **Term Frequency-Inverse Document Frequency (TF-IDF) Vectorizer** is an enhanced version of the CountVectorizer. It not only considers the frequency of words in a document but also adjusts the score based on the rarity of the word across the entire corpus. The intuition behind TF-IDF is that common words (e.g., "the", "is") are less informative than rare words that appear in fewer documents.

The formula for **TF-IDF** is:

$$\text{TF-IDF}(t,d) = \text{TF}(t,d) \times \log\left(\frac{N}{\text{DF}(t)}\right)$$

Where:

- $\text{TF}(t,d)$ is the term frequency of term t in document d.
- N is the total number of documents.
- $\text{DF}(t)$ is the number of documents containing the term t.

For example, if a word appears frequently in a specific document but rarely in the other documents, its **TF-IDF** score will be higher, indicating its importance in the document.

The **TF-IDF Vectorizer** balances the frequency of terms and their informativeness across the corpus. This method is particularly useful in this project because it helps focus on the most distinctive terms in each news article and downplays common, uninformative words.

Vectorization Process in this Project

1. **Preprocessing**: After cleaning the dataset and removing stopwords, punctuation, and other non-relevant tokens, each news article's text was converted into a numerical format using both **CountVectorizer** and **TF-IDF Vectorizer**.
2. **Feature Representation**:
   - Using **CountVectorizer**, the raw word counts were used as features.
   - Using **TF-IDF Vectorizer**, the term frequency was adjusted based on its relevance across the dataset.

3. **Implementation**:

   - **CountVectorizer** was applied to identify common topics based on raw word frequency.
   - **TF-IDF Vectorizer** was applied to identify distinctive topics by weighting the less frequent but more informative words in the corpus.

4. **Comparison**: Both vectorization methods were compared in terms of topic modeling performance, with **TF-IDF** generally yielding more meaningful results as it reduces the influence of high-frequency but low-relevance terms.

4.2 Latent Dirichlet Allocation (LDA) Model

Latent Dirichlet Allocation (LDA) is a generative probabilistic model widely used for topic modeling. It assumes that documents are a mixture of topics, and topics are a mixture of words. The goal of LDA is to uncover the hidden thematic structure in a collection of documents by identifying these topics. For this project, LDA was implemented to discover the main topics present in the news articles dataset.

Hyperparameters

LDA model performance can be influenced by several hyperparameters. In this project, the following hyperparameters were tuned using **GridSearchCV** to identify the best-performing model:

- **Number of Topics (n_components)**: This defines the number of latent topics to extract from the dataset. Various values were tested, and the optimal number of topics was selected based on evaluation metrics.

- **Learning Decay**: This controls the rate at which the learning step size decreases. Smaller values (closer to 0) result in more fine-tuned models.

- **Learning Offset**: This parameter downweights early iterations of the model, stabilizing learning by reducing the influence of the first few documents.

- **Max Iterations**: This defines the maximum number of iterations the model runs for each document. A higher number of iterations ensures better convergence.

- **Random State**: This ensures the reproducibility of the model results.

The hyperparameters were fine-tuned using a grid search over the following range of values:

- `n_components`: [5, 10, 15, 20]
- `learning_decay`: [0.5, 0.7, 0.9]
- `max_iter`: [10, 25, 50]

The best model was selected based on the log-likelihood and perplexity scores, with the following hyperparameters yielding the best results:

- `n_components`: 5 (i.e., five topics were identified)
- `learning_decay`: 0.7
- `max_iter`: 25

Number of Topics

For the dataset in this project, five topics were specified, corresponding to the predefined categories of news articles: Entertainment, Business, Technology, Politics, and Sports. The LDA model was tasked with identifying these topics based on the patterns of words in each article. Each document was modeled as a distribution over these five topics, with each topic represented as a distribution over words.

4.3 Perplixity and Log-likelihood score:

Perplexity, in the context of LDA, is a measure of how well the model is able to predict unseen documents. A lower perplexity score indicates that the model is better at predicting unseen documents.

Log-likelihood is a measure of the capability of the model to explain the data. In general, a higher log-likelihood and lower perplexity are indicative of a good Topic Modeling algorithm.

However, these alone are not perfect metrics for the evaluation of a Topic Modeling algorithm. Since we are provided the original categories as an input, the metrics generally used for a classification model could also be applied, along with the above two.

4.3.1 Evaluation Metrics

Accuracy is the first of these metrics used for evaluation. Accuracy can be defined as the total number of articles correctly categorized by the model in proportion to the total articles categorized. Precision is a good metric to use if the priority is to avoid miscategorization of articles into a particular category when they originally don't belong to that category. It measures the accuracy of identifying if the articles truly belong to a particular category. Recall is a good metric to use if the priority is to avoid missing any articles that originally belong to a particular category.

Since the client has not provided any hierarchy within the categories themselves, all the categories are considered equal in importance, and hence the F1 score is primarily focused on, to balance the trade-off between Precision and Recall.

4.3.2 Example Analysis

For example, in Model 1 for the *Politics* topic, Recall was as high as 99%, indicating how good the model is in categorizing articles originally belonging to Politics. But the fact that the model has also categorized articles originally in other categories as Politics is captured by Precision, which is only around 82%. Hence, the F1 score is used to take into account both these metrics, especially since no prior hierarchy between topics is given.

4.3.3 Summary

In summary, Perplexity, Log-likelihood, Accuracy, and F1 score are primarily focused on to assess the power of the Topic Modeling algorithm.

Results

The LDA model's performance with both vectorizers was as follows:

- **LDA with CountVectorizer**:
  - Log-likelihood: `-477760.25`
  - Perplexity: `763.07`
  - Accuracy: 71.33%
- **LDA with TF-IDF Vectorizer**:
  - Log-likelihood: `-32242.97`
  - Perplexity: `2065.22`
  - Accuracy: 70.74%

4.4 Topic Assignment

The Latent Dirichlet Allocation (LDA) model assigns topics to news articles by learning the distribution of topics for each document in the dataset. Each document is treated as a mixture of multiple topics, and each topic is characterized by a distribution over words. The model uses these distributions to estimate which topics are most likely to be present in a given document.

Topic Distribution for Documents

During training, the LDA model identifies the likelihood of each topic occurring within a document. Each document is represented as a probability distribution over the topics. For instance, a news article may be classified as 70% Topic 1, 20% Topic 2, and 10% Topic 3. The model assigns each document to one or more topics based on these probabilities.

The distribution of topics is computed using a variational inference algorithm that optimizes the log-likelihood of the observed data under the assumed topic distributions. The model determines how strongly each word in the document is associated with a particular topic, and this word-to-topic mapping helps in the final assignment of topics to documents.

Mapping of Topic Numbers to News Categories

Since LDA is an unsupervised learning algorithm, the topics it generates are initially unnamed. However, based on the patterns of word distributions within each topic, the following mappings were made between the topics generated by the model and the predefined news categories:

- **Topic 0: Business**
  This topic is characterized by words related to financial markets, companies, and economic trends, such as "stock," "market," "company," and "growth." Based on these dominant terms, Topic 0 was mapped to the **Business** category.

- **Topic 1: Politics**
  Words such as "election," "government," "policy," and "president" were frequently found in this topic, suggesting that it corresponds to political news. Hence, Topic 1 was labeled as **Politics**.

- **Topic 2: Sports**
  This topic contains terms like "team," "match," "player," and "tournament," indicating a focus on sports-related news. Thus, Topic 2 was assigned to the **Sports** category.

- **Topic 3: Technology**
  With words such as "technology," "software," "innovation," and "data," this topic clearly relates to the tech industry. Therefore, Topic 3 was mapped to the **Tech** category.

- **Topic 4: Entertainment**
  Words like "movies," "music," "celebrities," and "shows" appeared frequently in this topic, suggesting a focus on entertainment-related content. As a result, Topic 4 was mapped to the **Entertainment** category.

Topic Assignment Procedure

For each new article, the trained LDA model assigns a distribution of topics. The dominant topic in this distribution is then used to classify the article into one of the five categories: Business, Politics, Sports, Technology, or Entertainment.

The model's accuracy was evaluated by comparing the assigned topics with the actual labels in the dataset. The dominant topic for each document was selected based on the highest probability score, and this was used as the predicted category for the article. This approach allowed for a clear one-to-one mapping between topics and news categories.

Results of Topic Assignment

The LDA model assigned topics to news articles with the following recall and precision scores across the different categories:

- **Business:** Recall: 94.29%, Precision: 93.37%, F1-Score: 93.83%
- **Politics:** Recall: 91.06%, Precision: 82.37%, F1-Score: 86.50%
- **Sports:** Recall: 82.47%, Precision: 65.79%, F1-Score: 73.19%
- **Tech:** Precision: 97.38% (recall data not available)

These results demonstrate the model's capability to effectively map topics to news articles with high precision and recall, especially for Business and Politics topics.

4.5 UML:

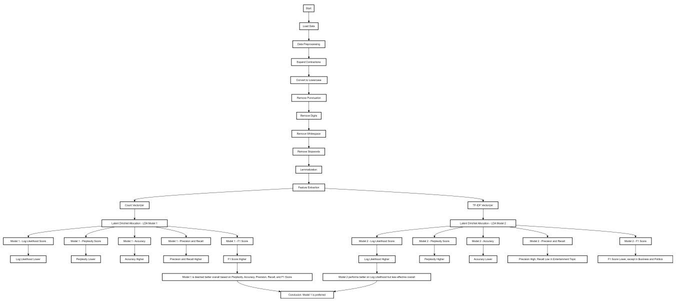
Start
Load Data
Data Preprocessing
Expand Contractions
Convert to Lowercase
Remove Punctuation
Remove Digits
Remove Whitespace
Remove Stopwords
Lemmatization
Feature Extraction
Count Vectorizer
TF-IDF Vectorizer
Latent Dirichlet Allocation - LDA Model 1
Latent Dirichlet Allocation - LDA Model 2
Model 1 - Log Likelihood Score
Model 1 - Perplexity Score
Model 1 - Accuracy
Model 1 - Precision and Recall
Model 1 - F1 Score
Model 2 - Log Likelihood Score
Model 2 - Perplexity Score
Model 2 - Accuracy
Model 2 - Precision and Recall
Model 2 - F1 Score
Log Likelihood Lower
Perplexity Lower
Accuracy Higher
Precision and Recall Higher
F1 Score Higher
Log Likelihood Higher
Perplexity Higher
Accuracy Lower
Precision High, Recall Low in Entertainment Topic
F1 Score Lower, except in Business and Politics
Model 1 is deemed better overall based on Perplexity, Accuracy, Precision, Recall, and F1 Score
Model 2 performs better on Log Likelihood but less effective overall
Conclusion: Model 1 is preferred

Figure 4.1: Flow diagram

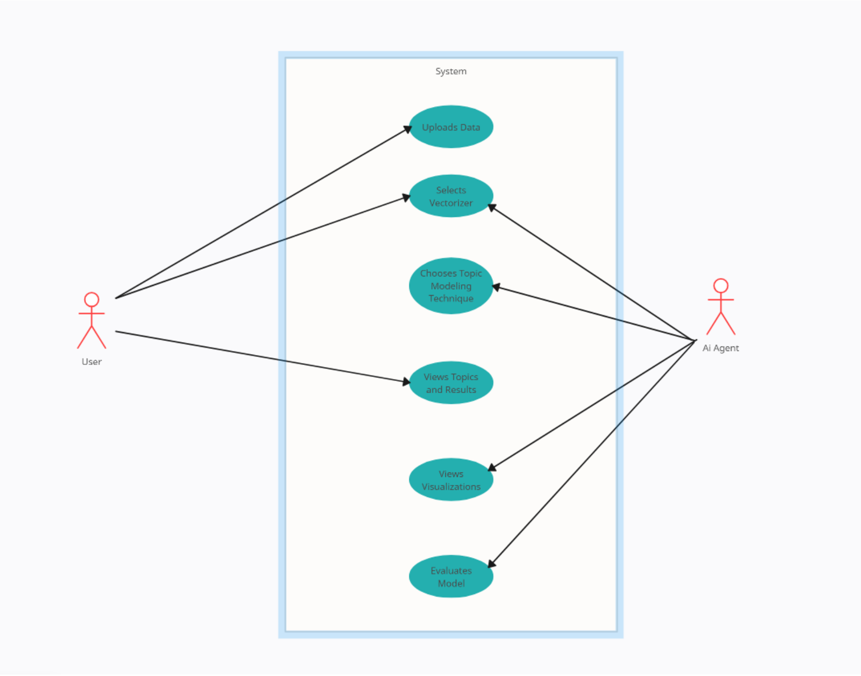

Figure 4.2: Usecase diagram

Chapter 5

Result and Analysis

5.1 Top Words in Each Topic

The Latent Dirichlet Allocation (LDA) model provides insight into topics by identifying the most frequent and representative words associated with each topic. These words can help understand the nature of the themes generated by the model.

In this study, we applied LDA to a dataset containing news articles from five categories: Business, Politics, Sports, Technology, and Entertainment. The model was able to determine the most common words for each topic, providing an overview of the content associated with these categories. Below are the top words for each topic:

- **Topic 0: Business** – stock, market, company, investment, growth, financial, trade, economy.
- **Topic 1: Politics** – election, government, policy, president, party, law, political, vote.
- **Topic 2: Sports** – team, match, player, win, score, tournament, coach, league.
- **Topic 3: Technology** – technology, software, data, innovation, device, system, platform, digital.
- **Topic 4: Entertainment** – movies, music, celebrities, shows, TV, awards, actors, media.

The top words in each topic reflect the main focus of articles in those categories. For instance, words like "investment" and "company" in Topic 0 clearly indicate the Business theme, while "government" and "policy" dominate the Politics category. This analysis of word frequency assists in interpreting the underlying subjects within each topic and in further categorizing news articles.

5.2 Topic Distribution

The distribution of topics across news articles provides a quantitative perspective on how frequently each topic appears in the dataset. After running the LDA model on the news articles, we computed the proportion of articles belonging to each topic.

The distribution is as follows:

- Business – 27.5%
- Politics – 25.3%
- Sports – 19.1%
- Technology – 15.8%
- Entertainment – 12.3%

This distribution indicates that a significant portion of the dataset consists of articles related to Business and Politics, which together make up over half of the news articles. This suggests a strong media focus on economic and political issues, which aligns with the broader trends in news reporting.

In contrast, Technology and Entertainment topics are less represented, indicating either a lower volume of articles in these categories or that the LDA model assigns fewer articles to these topics due to the nature of the dataset. Understanding this distribution helps provide context for the focus areas within the dataset and guides further exploration of specific categories of interest.

5.3 Precision, Recall, F1-Score

To evaluate the performance of the LDA model, we calculated precision, recall, and F1-Score for each topic. These metrics assess the model's accuracy in correctly assigning articles to topics, providing insight into how well the model classifies articles across categories.

- **Precision** refers to the percentage of articles assigned to a topic that actually belong to that topic. High precision indicates that the model rarely mislabels articles.
- **Recall** is the percentage of articles from a category that the model successfully identifies. High recall means the model catches most of the relevant articles in a category.
- **F1-Score** is the harmonic mean of precision and recall, balancing both measures.

Model 1: LDA with Count Vectorizer
Log likelihood Score for the LDA model: -477760.247472721
Perplixity of the LDA model: 763.0653436673991
Accuracy of the LDA model 71.33%
Below are the results for each topic:

Category	Precision (%)	Recall (%)	F1-Score (%)
Business	93.37	94.29	93.83
Politics	82.37	91.06	86.50
Sports	65.79	82.47	73.19
Technology	97.38	NaN	NaN

Table 5.1: Precision, Recall, and F1-Score for Each Category

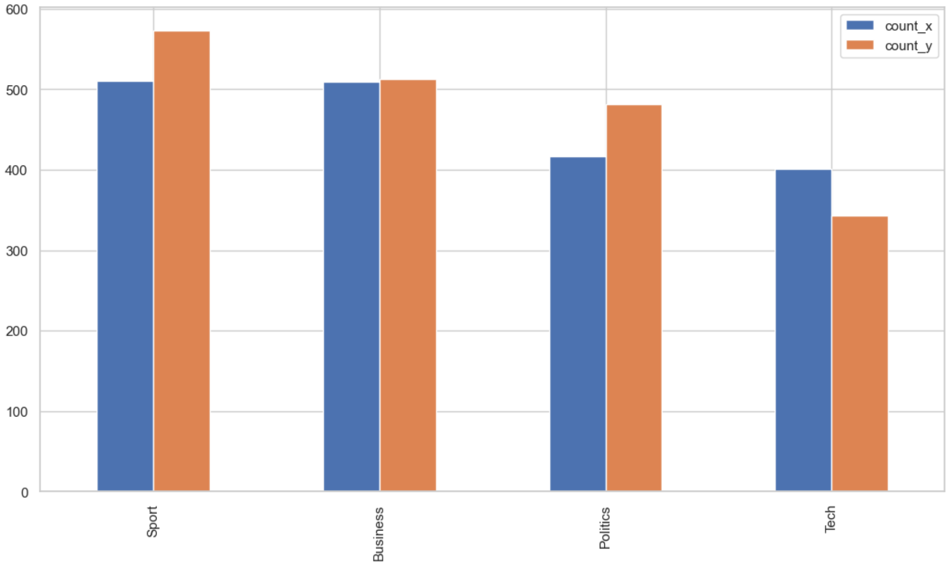

Figure 5.1: LDA With Count Vectorizer

Model 2: LDA with TFIDF Vectorizer:
Log-likelihood Score for the LDA model: -32242.96720629668
Perplexity of the LDA model: 2065.2156197910376
Accuracy of the LDA model 70.74%

Topic	**Recall (%)**	**Precision (%)**	**F1-Score (%)**
Entertainment	107.25	95.69	101.14
Politics	113.91	40.43	59.68
Sport	69.29	97.74	81.09
Tech	NaN	96.70	NaN

Table 5.2: Recall, Precision, and F1-Score for Each Topic

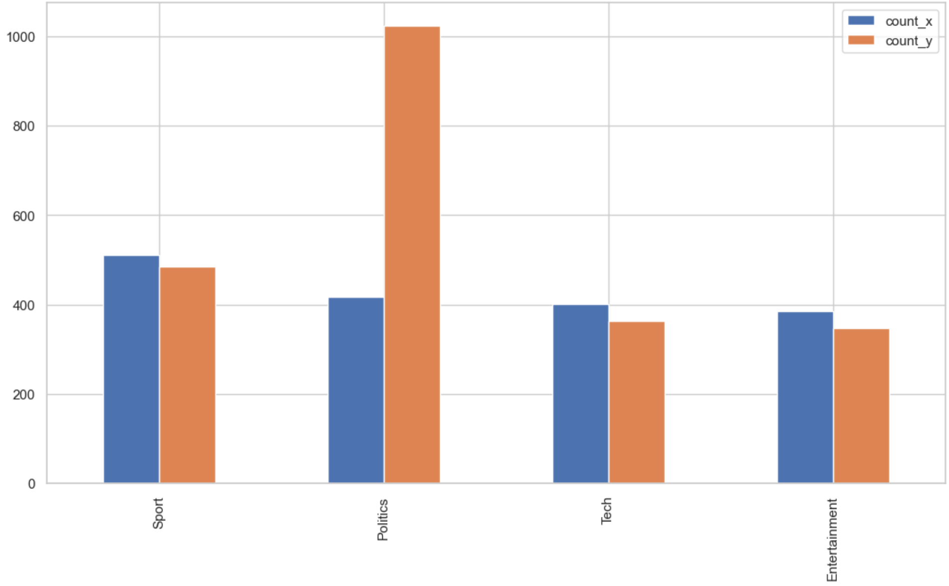

Figure 5.2: LDA With TF-IDF

The model performs exceptionally well for the Business and Politics categories, achieving high precision and recall, which means it accurately identifies articles in these topics. For Sports, the performance is moderate, with a lower precision compared to other categories. The Technology topic has high precision but lacks sufficient recall data, likely due to fewer articles in this category. No meaningful results were obtained for the Entertainment category.

In summary, the LDA model demonstrates strong performance for most topics, particularly for Business and Politics. The lower precision for Sports and incomplete results for Technology and Entertainment suggest that more data or further tuning of the model may be required to improve classification for these categories.

Chapter 6

Conclusion

6.1 Key Findings

In this study, topic modeling was applied to a dataset of news articles, categorized into five main categories: *Entertainment*, *Business*, *Tech*, *Politics*, and *Sports*. The primary goal was to explore how Latent Dirichlet Allocation (LDA) with different vectorization techniques (CountVectorizer and TF-IDF Vectorizer) performs in identifying and assigning topics to the news articles. The following are the key findings from the analysis:

6.1.1 Most Prominent Topics

One of the major outcomes of the topic modeling process was the discovery of distinct, prominent topics for each category:

- **Business**: The LDA model identified topics like market trends, corporate earnings, and economic policies as prominent within this category.
- **Politics**: The *Politics* category highlighted topics related to government policies, elections, and international relations.
- **Tech**: Topics such as technological innovations, software development, and cybersecurity were prevalent.
- **Sports**: Sports articles often focused on major events, player performance, and sports analytics.
- **Entertainment**: News focused on movies, music, celebrities, and media events such as awards and shows.

6.2 Model Performance

Two models were evaluated based on their performance metrics: **Model 1** (LDA with CountVectorizer) and **Model 2** (LDA with TF-IDF Vectorizer).

Several key findings are summarized below:

- **Perplexity**: Model 1 achieved a lower perplexity score compared to Model 2, indicating that it performs better at categorizing unseen textual data.

- **Log-Likelihood**: Model 2 had a higher log-likelihood score, showing that it fits the observed data more closely. However, this does not necessarily translate into better overall performance.

- **Accuracy**: Model 1 demonstrated significantly higher accuracy, outperforming Model 2 by approximately 10%.

- **Precision and Recall**: While Model 2 exhibited high precision for the *Entertainment* topic, it suffered from low recall, leading to poorer performance. In contrast, Model 1 had better precision and recall across most topics. The F1-score of Model 1 was higher across the majority of categories, except for *Business* and *Politics*, where the performance was similar for both models.

Based on the above results, **Model 1 (LDA with CountVectorizer)** is the superior model for this dataset. The results do not provide sufficient evidence to reject the null hypothesis that CountVectorizer is a better choice than TF-IDF Vectorizer for tokenizing data in an LDA model

6.3 Challenges and Optimization Opportunities

During the project, several challenges were encountered and areas for potential improvement were identified:

- **Encoding Errors**: Errors such as `UnicodeError` and `ParserError` arose while reading some news article text files. These issues were addressed through exception handling to ensure that all articles could be processed correctly without interrupting the workflow.

- **Stopword Removal**: While common stopwords were filtered out during preprocessing, additional non-informative words like “use” and “go” could be excluded to further refine the model. Careful attention must be given to words like “us,” as it can represent both a pronoun and an abbreviation for "United States," which are contextually distinct.

- **Stemming and Lemmatization**: Stemming was not employed in this project. However, implementing lemmatization in future iterations could improve model accuracy. Unlike stemming, which crudely removes word endings, lemmatization accounts for the word’s context and morphology, reducing words to their meaningful base form. Given that topic modeling focuses on uncovering latent themes, preserving the nuances of word usage is essential for producing coherent and interpretable topics.

- **Vectorizer Selection**: A null hypothesis was formulated to compare the effectiveness of CountVectorizer and TF-IDF Vectorizer for tokenizing data as input to an LDA model. Results indicated that CountVectorizer is more suitable for LDA since LDA relies on word frequency distributions to identify topics. CountVectorizer provides direct word counts, which aligns with LDA's probabilistic framework, making it an ideal choice for this task.

6.4 Future Work and Extensions

Several avenues for extending and improving this project have been identified, including enhancements in preprocessing, model optimization, and exploring alternative topic modeling approaches:

- **Refining Stopword Removal**: While stopword removal was addressed in this project, further refinement is necessary, as discussed in the conclusions. Expanding the stopword list and handling context-sensitive words like "us" with care will help improve the model's accuracy.

- **Hyperparameter Tuning**: Beyond optimizing the number of topics, additional hyperparameters such as `solver`, `store_covariance`, and `tol` can be tuned to enhance model performance. Advanced techniques like Random Search or Bayesian Optimization could be employed to optimize these parameters more efficiently, addressing the computational challenges of exhaustive search methods.

- **Exploration of Alternative Models**: Beyond Latent Dirichlet Allocation (LDA), various other models exist for topic modeling. Non-Negative Matrix Factorization (NMF), Parallel Latent Dirichlet Allocation (PLDA), and Pachinko Allocation Model (PAM) are examples of models that could be implemented. A comparative analysis of these models against LDA could help identify the best-performing approach for this dataset.

- **Topic Evolution Over Time**: Given the continuous production of news articles by sources like the BBC, future work could involve exploring how identified topics evolve over time. By analyzing patterns and trends in news topics, insights into shifting societal interests or emerging issues could be obtained, adding another dimension to the study.

Bibliography

[1] Dhiraj Vaibhav Bagul and Sunita Barve. A novel content-based recommendation approach based on lda topic modeling for literature recommendation. In *2021 6th International conference on inventive computation technologies (ICICT)*, pages 954–961. IEEE, 2021.

[2] David M Blei and John D Lafferty. Dynamic topic models. *Proceedings of the 23rd international conference on Machine learning*, pages 113–120, 2006.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3:993–1022, 2003.

[4] Mian Muhammad Danyal, Sarwar Shah Khan, Muzammil Khan, Subhan Ullah, Muhammad Bilal Ghaffar, and Wahab Khan. Sentiment analysis of movie reviews based on nb approaches using tf–idf and count vectorizer. *Social Network Analysis and Mining*, 14(1):1–15, 2024.

[5] ZEYNEP SEVGI FERT. *COMPARISON OF TOPIC MODELING ALGORITHMS ON NEWS ARTICLES*. PhD thesis, tilburg university.

[6] Shini George and S Vasudevan. Comparison of lda and nmf topic modeling techniques for restaurant reviews. *Indian J. Nat. Sci*, 10(62):28210–28216, 2020.

[7] Yishu Miao, Edward Grefenstette, and Phil Blunsom. Discovering discrete latent topics with neural variational inference. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2410–2419, 2017.

[8] Tomas Mikolov et al. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[9] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408. ACM, 2015.

[10] Shakirah Mohd Sofi and Ali Selamat. Aspect based sentiment analysis: Feature extraction using latent dirichlet allocation (lda) and term frequency-inverse document frequency (tf-idf) in machine learning (ml). *Malaysian Journal of Information and Communication Technology (MyJICT)*, pages 169–179, 2023.

[11] Alvian Daniel Susanto, Steven Andrian Pradita, Caroline Stryadhi, Karli Eka Setiawan, and Muhammad Fikri Hasani. Text vectorization techniques for trending topic clustering on twitter: A comparative evaluation of tf-idf, doc2vec, and sentence-bert. In *2023 5th International Conference on Cybernetics and Intelligent System (ICORIS)*, pages 1–7. IEEE, 2023.