Topic Modelling using Latent Dirichlet Allocation (LDA)

Shubham Gupta (202318052) *MSc Data Science* *DAIICT, Gandhinagar*

Digesh Patel (202318038) *MSc Data Science* *DAIICT, Gandhinagar*

Mayan Bhut (202318043) *MSc Data Science* *DAIICT, Gandhinagar*

Yash Chaudhary (202318022) *MSc Data Science* *DAIICT, Gandhinagar*

ABSTRACT

The objective of this study was to uncover hidden topics within a collection of BBC News articles using Natural Language Processing (NLP) and clustering techniques. The dataset comprised 2225 articles across five topics: business, entertainment, politics, sports, and technology, with duplicates removed. Initial analysis revealed that business and sports topics dominated the corpus, with varying article lengths among topics. Further exploration exposed frequent occurrence of stop-words and short-length words, prompting data cleaning including stop-word removal, lemmatization, and eliminating unnecessary elements. Vectorization with TF-IDF was followed by clustering using Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA).

LDA emerged as the superior algorithm, effectively identifying and clustering articles into distinct topics, as evidenced by visualization techniques such as pyLDA plots and word clouds. LSA also identified five topics but exhibited poorer clustering performance. Thus, LDA was deemed the optimal choice for BBC News article topic modeling.

INTRODUCTION

Topic modeling is a strong tool for identifying underlying topics in a corpus of text data. This approach, which makes use of Natural Language Processing (NLP) and clustering algorithms, makes it easier to identify and extract hidden ideas or themes inherent inside a collection of texts. In the context of this study, the dataset consists of a corpus of news stories taken from BBC News and divided into five categories: business, entertainment, politics, sport, and technology. The primary goal is to create an unsupervised machine learning model capable of extracting these latent topics from news articles.

Data preparation is an important first step in topic modeling BBC News stories. This entails gathering text documents from multiple themes and combining them into a single dataset while removing duplicate articles. This consolidation simplifies subsequent analysis by offering a cohesive corpus to work with.

Following data aggregation, the dataset is explored to get insights into both the substance of the news stories and the distribution of topics across the corpus. This investigation step assists analysts in understanding the common topics in the dataset and how they are spread across different sections or time periods.

It's time to clean up the dataset after it has been examined in order to improve its quality and suitability for modeling. This entails streamlining the text and eliminating superfluous or unnecessary details. To make the text more streamlined, strategies such eliminating stop words, short words, special characters, numerals, and excess white space are used.

Lemmatization is also used to reduce words to their most basic form, which makes analysis and interpretation more efficient. Analysts may guarantee uniformity throughout the dataset by standardizing the text in this way, which facilitates topic identification and analysis.

Term Frequency-Inverse Document Frequency (TF-IDF) vectorization is the next stage after preprocessing and cleaning the data. The textual input is converted into numerical vectors by this process, which is necessary in order to apply clustering methods for topic modeling. TF-IDF produces a numerical representation of the text that may be input into machine learning models by calculating the relevance of each word in a document in relation to its frequency over the whole corpus.

Finally, modeling is carried out using clustering methods like Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA). These algorithms evaluate vectorized data to find patterns and cluster papers based on thematic similarities. This study uses approaches to identify hidden subjects in BBC News items, providing useful insights on their substance and themes.

MATHEMATICAL FORMULATION

In this section we will discuss the two numerical methods namely:

- Latent Dirichlet Allocation (LDA)
- Latent Semantic Analysis (LSA)

1) Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a popular generative statistical model used for topic modeling, which is a technique for discovering abstract topics in a collection of documents. Here's a detailed explanation of the mathematical formulation of LDA:

- *Assumptions*
    - LDA assumes that documents are generated in the following way: first, a set of topics is chosen for the entire corpus, and then for each document, a mixture of these topics is sampled. Finally, words are generated based on this mixture of topics.
    - Each document exhibits a blend of multiple topics.
    - Each topic is characterized by a distribution of words. Topics capture the semantic themes present in the documents.
- *Notations*
    - *D*: The number of documents in the corpus.
    - *N*: The number of words in a document.
    - *K*: The number of topics.
    - *V*: The vocabulary size (total number of unique words in the corpus).
    - *w*: A specific word in the vocabulary.
    - *θd*: The topic distribution for document *d*. It represents the proportion of each topic in document d.
    - *βk*: The word distribution for topic *k*. It represents the probability of each word occurring in topic *k*.
- *Generative Process*

  For each document d:

  Sample a topic distribution θd from a Dirichlet distribution with parameter α. The Dirichlet distribution ensures that the topic distribution is sparse.

  For each word n in document d:

  Sample a topic, z d,n from the multinomial distribution defined by θd. This determines which topic generates the current word.

  Sample a word w d,n from the multinomial distribution defined by β z d,n. This selects the specific word from the chosen topic.
- *Probabilistic notations*
    - p(θd ∣ α): Probability of the topic distribution for document d given the Dirichlet parameter α.
    - p(z d,n ∣ θd): Probability of topic z d,n given the topic distribution θd.
    - p(w d,n ∣ z d,n ,β): Probability of word w d,n given the topic z d,n and word distribution β.
- *Likelihood*
    - The likelihood of observing the entire corpus is calculated as the product of the probabilities of generating each word in each document, integrated over all possible topic distributions θd.
    - This integral is usually analytically intractable, so approximate inference methods such as variational inference or Gibbs sampling are employed.

- *Inference*
    - Given a corpus, the inference process involves estimating the parameters (topic distributions θd and word distributions β k) that maximize the likelihood of observing the corpus.
    - This typically involves iterative algorithms like variational inference or Gibbs sampling to approximate the posterior distribution over latent variables.
- *Parameter Estimation*
    - Once the topic distributions θd and word distributions βk are inferred, they are used to:
    - Identify the most probable topics for each document.
    - Identify the most probable words associated with each topic.
- *Hyperparameters*
    - α and β are hyperparameters of the model that control the sparsity of topic distributions and word distributions, respectively.
    - Proper selection of hyperparameters is crucial for the model's performance.

2) Latent Semantic Analysis (LSA)

Latent Semantic Analysis (LSA) is a technique used for dimensionality reduction and capturing the latent semantic structure in a collection of documents. Here's an explanation of its mathematical formulation:

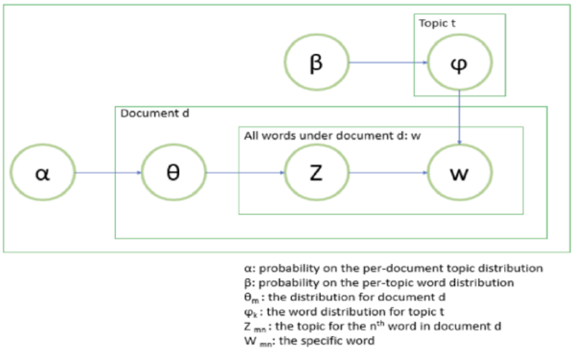
- *Term-Document Matrix:*
    - Given a corpus of D documents and a vocabulary of V unique terms, construct a term-document matrix A of size V×D, where each element a ij represents the frequency of term i in document j.
    - Optionally, normalize the term frequencies using techniques like TF-IDF (Term Frequency-Inverse Document Frequency) to downweight frequently occurring terms.
- *Singular Value Decomposition (SVD):*
    - Apply Singular Value Decomposition to the term-document matrix A. SVD decomposes A into three matrices: U, Σ, and VT
    - Let A be of size V×D, then SVD decomposes it as A=UΣVT, where:

      U is an V×r orthogonal matrix, representing the term space.
      Σ is an r×r diagonal matrix, where r is the rank (number of singular values) of A. It contains the singular values in descending order.
      Vt is a r×D orthogonal matrix, representing the document space.
- *Dimensionality Reduction:*
    - Truncate the matrices U, Σ, and VT to reduce the dimensionality of the representation.
    - Retain only the top k singular values and their corresponding columns in U, rows in Σ, and rows in VT where k is the desired dimensionality of the latent space.
- *Semantic Space Representation:*
    - The reduced matrices Uk, Σk, and VkT represent the documents and terms in a lower-dimensional semantic space.
    - The i-th column of UK represents the i-th term's vector representation in the semantic space.
    - The i-th row of VkT represents the i-th document's vector representation in the semantic space.
- *Document Similarity:*

  To measure similarity between documents, compute the cosine similarity between their vector representations in the latent semantic space.
- *Term Similarity:*

  To measure similarity between terms, compute the cosine similarity between their vector representations in the latent semantic space.

METHODOLOGY AND IMPLEMENTATION


Topic t
β
φ
Document d
All words under document d: w
α
θ
Z
w
α: probability on the per-document topic distribution
β: probability on the per-topic word distribution
θm : the distribution for document d
φk : the word distribution for topic t
Z mn : the topic for the nth word in document d
W mn: the specific word

Latent Dirichlet Allocation:

LDA is an unsupervised learning technique an its typical example is Topic Modelling. The assumption is that each document mix with various topics and every topic mix with various words. The below figure shows the flow of LDA algorithm:

For a particular document d, we get a topic distribution which is θd. From this distribution topic t will be chosen and selecting corresponding word from phi.

Steps:

The methodology for Topic Modeling on BBC News Articles begins with data pretreatment and investigation to determine the dataset's properties. Initially, the dataset contains 2225 news stories organized into five categories: business, entertainment, politics, sports, and technology. Duplicate articles are discovered and eliminated, leaving a dataset of 2127 unique news pieces.

Exploratory investigation suggests that business and sports receive a higher share of news stories than other topics. Furthermore, articles on business topics are typically shorter in length, whereas articles on politics and entertainment are longer, with the majority of articles lasting close to 500 words.

More research includes examining the frequency of individual words (unigrams), two-word sequences (bigrams), and three-word sequences (trigrams). The top 20 most frequent words are primarily stop-words and short-length words, indicating the need for data cleaning. Stop-words, short-length words, special characters, digits, unnecessary whitespace, and newline characters are all removed throughout the cleaning process. Furthermore, lemmatization is used to reduce each word to its root form, making grouping easier during modeling.

Following cleaning, the dataset is vectorized using the Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer, which weights words based on their frequency in documents relative to the overall corpus.

$$tf(t,d) = \ count\ of\ t\ in\ d\ /\ number\ of\ words\ in\ d$$

$$tf - idf(t,d)\ =\ tf(t,d)\ *\ idf(t)$$

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y
df_x = number of documents containing x
N = total number of documents

Modeling is done with two clustering algorithms: Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA).

- LDA is a generative probabilistic model that assumes each document is a mix of themes, with each topic being a distribution of words. It finds latent topics in the corpus and assigns a distribution to each document based on those subjects. The LDA algorithm increases the likelihood of observing the corpus by iteratively updating topic and word distributions.
- LSA is a dimensionality reduction approach that uses Singular Value Decomposition (SVD). It encodes texts and concepts in a lower-dimensional semantic space while capturing latent semantic links. However, because of its linear algebraic method, LSA may do less well than LDA in clustering documents into subjects.

Visualization tools like as pyLDA plots and word clouds are used to evaluate LDA's performance in detecting subjects and clustering news items. The LSA scatter plot is also evaluated in order to compare its performance to that of LDA. It is concluded that LDA is the best clustering algorithm for Topic Modeling on

BBC News Articles because it successfully recognizes and groups news articles into separate subjects.

SOLUTION TO THE PROBLEM

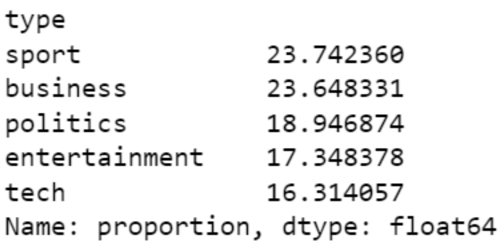
The methodology for Topic Modeling on BBC News Articles begins with data preprocessing to eliminate duplicates and better understand the dataset's properties. Vectorization with TF-IDF is used to convert cleaned text input into numerical form for modeling.

Two clustering methods, Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA), are then used to detect and extract hidden topics from news articles. LDA, a probabilistic model, is effective in identifying topics and clustering documents by leveraging the underlying distributions of words and subjects. Visualization approaches such as the pyLDA plot and word clouds demonstrate LDA's capacity to discriminate between subjects. In contrast, LSA, which is based on linear algebraic principles, fails to accurately cluster articles into relevant subjects.

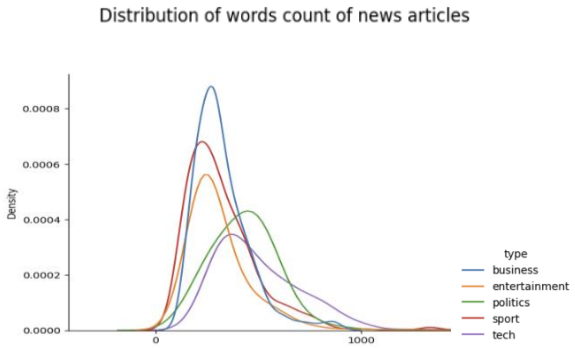
As a result, LDA emerges as the preferred clustering algorithm, offering a robust solution for Topic Modeling on BBC News Articles by effectively organizing them into distinct thematic categories.

RESULTS / OBSERVATIONS

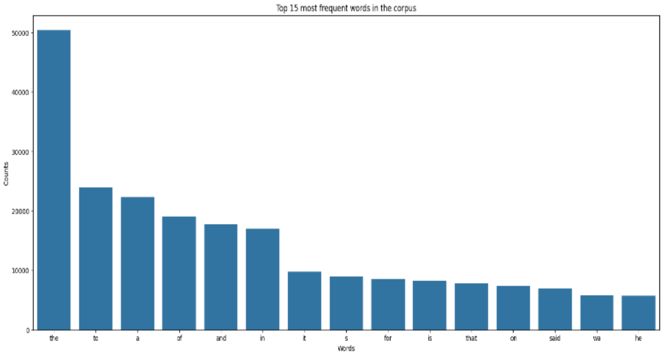
- The dataset consist of 2225 rows and 3 columns (unique_id,news,type)
- The news articles are of 5 unique types.
- The dataset has no null values but it has 98 duplicate news articles.
- The new shape of the dataset is 2127 rows with 3 columns after removal of duplicates



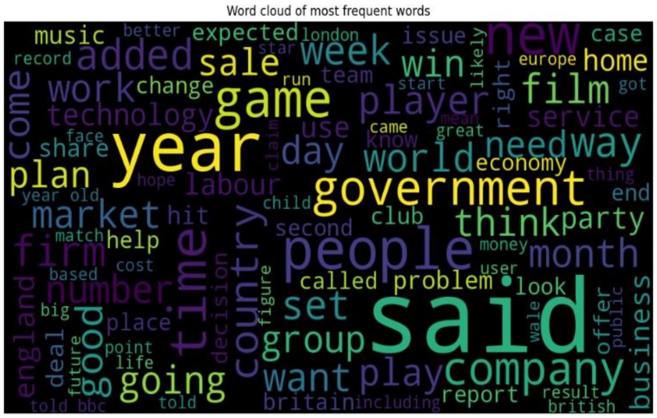
```
type
sport            23.742360
business         23.648331
politics         18.946874
entertainment    17.348378
tech             16.314057
Name: proportion, dtype: float64
```



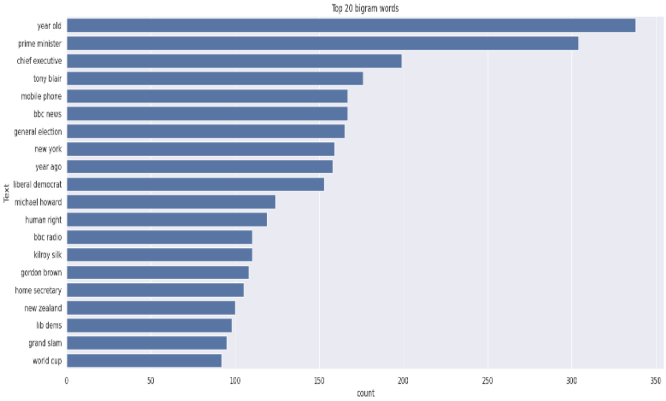
Distribution of words count of news articles



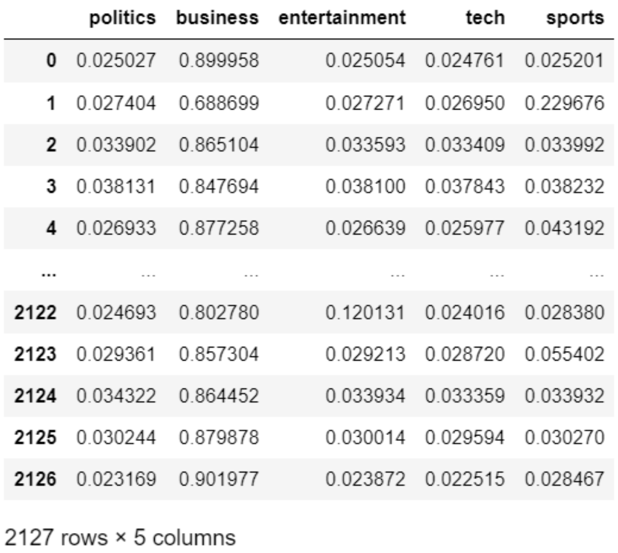
Top 15 most frequent words in the corpus



Word cloud of most frequent words



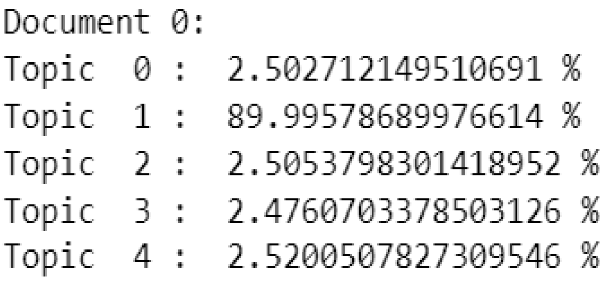
Top 20 bigram words



	politics	business	entertainment	tech	sports
0	0.025027	0.899958	0.025054	0.024761	0.025201
1	0.027404	0.688699	0.027271	0.026950	0.229676
2	0.033902	0.865104	0.033593	0.033409	0.033992
3	0.038131	0.847694	0.038100	0.037843	0.038232
4	0.026933	0.877258	0.026639	0.025977	0.043192
...	...	...	...	...	...
2122	0.024693	0.802780	0.120131	0.024016	0.028380
2123	0.029361	0.857304	0.029213	0.028720	0.055402
2124	0.034322	0.864452	0.033934	0.033359	0.033932
2125	0.030244	0.879878	0.030014	0.029594	0.030270
2126	0.023169	0.901977	0.023872	0.022515	0.028467

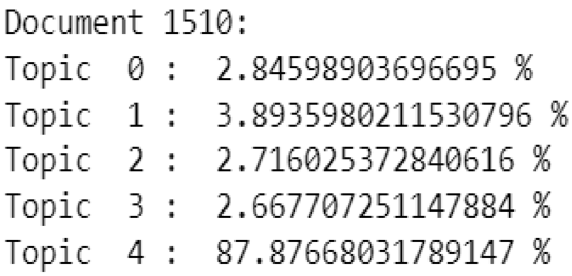
2127 rows × 5 columns

- ➢ The above figure shows the scores (how much a document belongs to a given class) of each document belonging to particular topic.



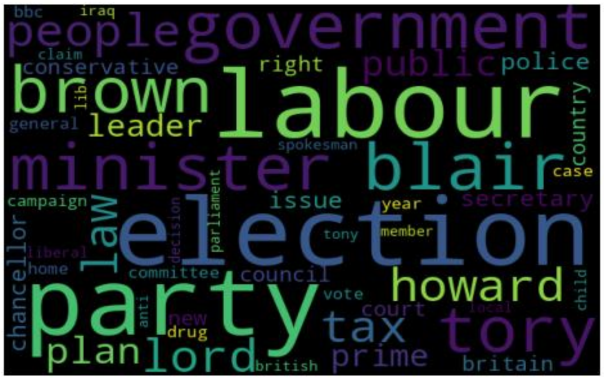
```
Document 0:
Topic  0 :  2.502712149510691 %
Topic  1 :  89.99578689976614 %
Topic  2 :  2.5053798301418952 %
Topic  3 :  2.4760703378503126 %
Topic  4 :  2.5200507827309546 %
```

- ➢ From above output we can see that, The first document is more belonging to the Topic 1 (business).



```
Document 1510:
Topic  0 :  2.84598903696695 %
Topic  1 :  3.8935980211530796 %
Topic  2 :  2.716025372840616 %
Topic  3 :  2.667707251147884 %
Topic  4 :  87.87668031789147 %
```

- ➢ From above output we can see that, The 1510th document is more belonging to the Topic 4 (sports).



Word cloud for topic (POLITICS)

CONCLUSION

The topic modeling approach starts with data preprocessing, which identifies and removes duplicates, yielding a dataset of 2127 news articles. Exploration finds that article proportions vary by topic, with business and sports dominating. The dataset is further cleaned by deleting stop-words and short-length words, which reduces it by 50%. Vectorization with TF-IDF is used, followed by modeling with LDA and LSA. LDA effectively clusters articles into five categories, outperforming LSA and demonstrating its applicability for BBC News Article Topic Modeling.

REFERENCES

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, null (3/1/2003), 993–1022.

[2] M. Abramowitz and I. Stegun, editors. *Handbook of Mathematical Functions*. Dover, New York, 1970.

[3] B. de Finetti. *Theory of probability. Vol. 1-2*. John Wiley & Sons Ltd., Chichester, 1990. Reprint of the 1975 translation.

[4] J. Dickey. Multiple hypergeometric functions: Probabilistic interpretations and statistical uses. *Journal of the American Statistical Association*, 78: 628-637, 1983.

[5] G. Ronning. Maximum likelihood estimation of Dirichlet distributions. *Journal of Statistcal Computation and Simulation*, 34(4): 215-221, 1989.