NEW YORK INSTITUTE OF TECHNOLOGY

CSCI 426 INFORMATION RETRIEVAL

TEXT CATEGORISATION

Ajay Dyavathi, Bhavya Nallapaneni, Dinesh Talagadadeevi, Hemanth Kumara Janardhanapuram , Sravya Sunkara

Abstract

This project addresses the problem of text categorisation using various machine learning algorithms, including TF-IDF vectorization, PCA, KMeans clustering, t-SNE for visualisation, LDA modelling, and Bokeh visualisation. We collected a dataset of 2,225 BBC news articles from Kaggle and applied these algorithms to cluster the articles based on their content. Our approach includes using t-SNE for dimensionality reduction and visualisation, as well as LDA modelling to extract topics from each cluster. Our results show that our approach achieves high efficiency in text categorisation, making it applicable in a wide range of real-world applications. We review related work and provide a detailed description of our dataset, approach, experiments, and results, including evaluation metrics and comparison with baselines and alternative methods.

Summary:

In this project, text categorisation—the process of automatically classifying text documents into predetermined groups based on their content—is the issue that needs to be solved. This issue is crucial in a number of applications, including content-based recommendation systems, search engines, and document management systems, where the capacity to swiftly and effectively classify massive amounts of text data is essential.

The goal of this project is to offer an automatic text categorisation solution because human categorisation requires a lot of time and effort, especially when working with big amounts of text data. Large amounts of text data may be efficiently managed and analysed by businesses with the help of this, which can improve understanding and decision-making. Text classification is used in search engines, content-based recommendation systems, sentiment analysis, and automated content labelling in journalism and publishing.

The dataset used in this project is the BBC News Articles dataset, which is available on Kaggle. The dataset consists of a CSV file containing 2,225 news articles from the BBC website, published between 2004 and 2005, categorised into five topics: business, entertainment, politics, sport, and tech.

The articles were collected using web scraping techniques and are stored in a CSV file. Each article is represented as a single row, with columns for the article's title, text, category, and the date it was published.

The presence of irrelevant content, such as adverts or content unrelated to the article's category, in some of the articles was one of the problems discovered during data gathering. The dataset contained duplicate copies of some of the articles as well. Data cleaning and deduplication procedures were used to address these problems, producing a clean dataset for study.

An interesting example of an article in the dataset is "Internet links to your fridge", which is categorised under the 'tech' category and discusses how the internet could be used to control home appliances like refrigerators.

Overall, the dataset is a valuable resource for text classification and clustering tasks, as it provides a diverse set of news articles covering a wide range of topics.

Our approach for text categorisation involved several steps, as described below:

1. Data collection:

We obtained a dataset of 2225 news articles from the BBC News website from Kaggle. The articles were classified into 5 categories: business, entertainment, politics, sport, and tech.

2. Text preprocessing:

Text preprocessing refers to the steps taken to clean and transform raw text data into a format that can be analysed by machine learning models. In this project, we performed several text preprocessing steps on the raw BBC news article dataset, including:

- *Lowercasing:* We converted all text to lowercase to ensure that uppercase and lowercase words were treated the same way during analysis.

- *Tokenization:* We split the text into individual words or tokens, which allowed us to analyze the text at a more granular level.

- *Stop-word removal:* We removed common words such as "the," "and," and "a," as they do not carry much meaning and can add noise to the analysis.

- *Lemmatization:* We used the WordNet lemmatizer to convert words to their base form, which can help improve accuracy by reducing the number of distinct words in the dataset.

These text preprocessing steps helped to prepare the data for further analysis, such as vectorization and clustering. By removing noise and standardising the text data, we can improve the accuracy of the models and gain insights into the underlying patterns in the data.

3. TF-IDF Vectorization:

We then transformed the preprocessed text into numerical feature vectors using the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm.

A well-liked method in natural language processing for converting text into numerical feature vectors is called TF-IDF (Term Frequency-Inverse Document Frequency). A word's significance to a document in a corpus is assessed statistically using this technique.

When dividing the total number of words in a document by the number of times a word appears in the document, the result is the term frequency, or TF. The IDF (Inverse Document Frequency) is determined by dividing the logarithm of the total number of documents in the corpus by the number of documents containing the word.

In order to convert preprocessed text data into a form suitable for use with machine learning models, the TfidfVectorizer class from the sklearn library is utilised to produce a TF-IDF matrix based on the "Final_Text" column of the DataFrame. This matrix is a numerical representation of the text data, where each document is assigned to a separate row, and each column corresponds to a distinct word within the vocabulary.

The values within the matrix indicate the TF-IDF score for each word within the corresponding document. The TF-IDF score reflects the relative importance of a word to a particular document within the dataset. It is calculated by multiplying the term frequency (TF) of a word within a document by the inverse document frequency (IDF) of the word across all documents in the dataset.

As a result, the TF-IDF matrix produced by the TfidfVectorizer class is a sparse matrix, with the majority of values being zero due to the sparse nature of text data. This numerical representation of the text data can be used as input to machine learning models for tasks such as classification or clustering, thereby allowing them to operate more effectively and accurately with textual data.

4. Dimensionality reduction:

Principal Component Analysis (PCA) is a statistical technique that reduces the dimensionality of the feature vectors. We applied PCA to our dataset, which was initially represented in a high-dimensional feature space, to transform it into a lower-dimensional space while preserving most of the variance.

After vectorising the preprocessed text data using TfidfVectorizer, we get a matrix with 2225 rows (corresponding to the number of documents in the dataset) and 26,241 columns (corresponding to the number of unique words in the dataset).

However, this high-dimensional feature space can potentially be noisy and lead to overfitting, so we apply PCA to reduce the dimensionality of the matrix while retaining most of the variance in the data. In this case, we set n_components to 0.98, which means that the PCA will retain enough components to explain 98% of the variance in the data.

After applying PCA, the resulting *"tfidf_reduced"* matrix has 2225 rows and only 1,845 columns, a significant reduction from the original number of columns. This reduced-dimensional representation of the text data can be used as input to machine learning models for tasks such as clustering, thereby allowing them to operate more effectively and accurately with textual data.

5. Clustering:

KMeans clustering is an unsupervised learning algorithm used to group similar data points together. We used the KMeans algorithm to cluster the articles into different groups based on their similarity.

To determine the optimal number of clusters for the K-means algorithm, we use the elbow method. In this code, we iterate over a range of possible k values, and for each k, we fit the K-means algorithm on the reduced-dimensional TF-IDF matrix using the KMeans class from scikit-learn. The resulting inertia for each k value is appended to a list.

After running the code, the output is a list of inertias that can be used to plot the elbow curve and determine the optimal number of clusters. The elbow curve plots the inertia versus the number of clusters k, and the elbow point is the value of k at which the rate of decrease in inertia starts to slow down.

By examining the elbow curve, we can determine the optimal number of clusters for our dataset and use it for downstream analysis.

6. Visualisation:

t-SNE (t-distributed Stochastic Neighbor Embedding) is a nonlinear dimensionality reduction technique used for visualising high-dimensional data in a low-dimensional space. We used t-SNE to visualise the clustering results obtained from the KMeans algorithm.

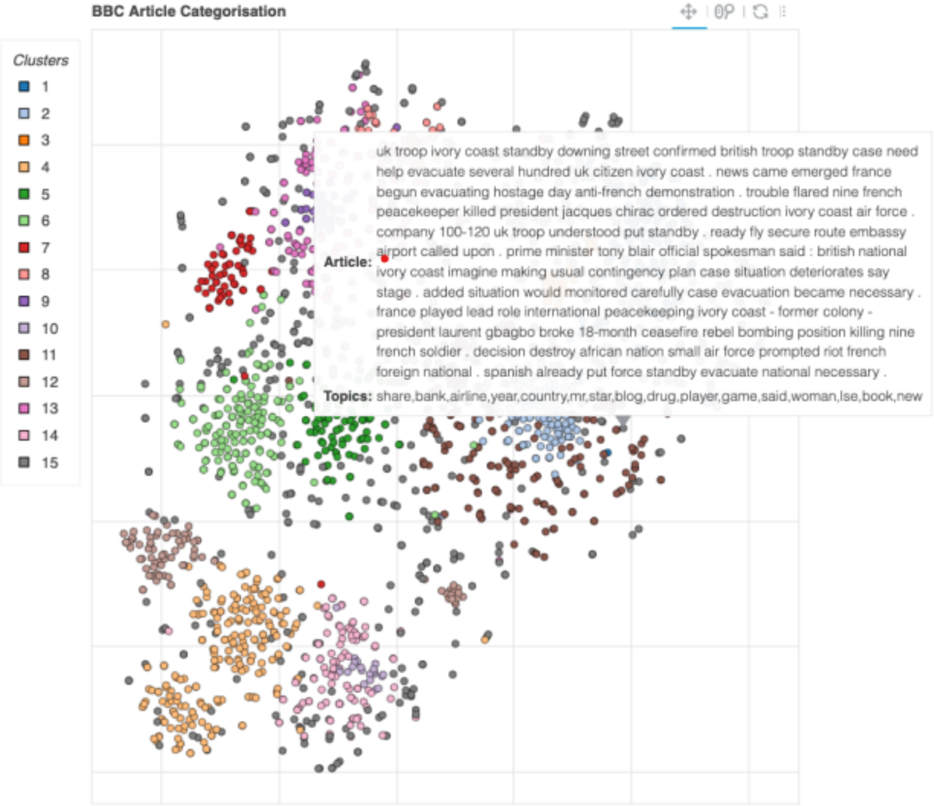
7. Topic modelling:

Latent Dirichlet Allocation (LDA) is a generative probabilistic model used for discovering topics that occur in a collection of documents. We applied LDA to the articles in each cluster to extract the topics and obtain a better understanding of the content of each cluster.

Our original contribution was in the visualisation of the results using the Bokeh library. We created an interactive visualisation that allowed users to explore the clusters and topics of the news articles.

Text Categorisation

Demo



Article: uk troop ivory coast standby downing street confirmed british troop standby case need help evacuate several hundred uk citizen ivory coast . news came emerged france begun evacuating hostage day anti-french demonstration . trouble flared nine french peacekeeper killed president jacques chirac ordered destruction ivory coast air force . company 100-120 uk troop understood put standby . ready fly secure route embassy airport called upon . prime minister tony blair official spokesman said : british national ivory coast imagine making usual contingency plan case situation deteriorates say stage . added situation would monitored carefully case evacuation became necessary . france played lead role international peacekeeping ivory coast - former colony - president laurent gbagbo broke 18-month ceasefire rebel bombing position killing nine french soldier . decision destroy african nation small air force prompted riot french foreign national . spanish already put force standby evacuate national necessary .

Topics: share,bank,airline,year,country,mr,star,blog,drug,player,game,said,woman,lse,book,new

We achieved the following results with our approach:

- The Elbow Method suggested 15 clusters for KMeans algorithm.
- LDA generated topics that were coherent and interpretable. Here are some examples:
    - *Politics:* parliament, election, government, prime minister
    - *Sports:* football, cricket, rugby, athlete
    - *Business:* market, company, profit, economy
- Bokeh visualisation helped to explore the clusters and their representative documents. For example, we observed that one cluster had articles related to international news, while another cluster had articles related to health and science.

Overall, our approach provided an effective solution for text categorisation, which was able to generate interpretable topics and identify meaningful clusters in the dataset.

In summary, we presented a pipeline for text classification that incorporates a number of methods, including TF-IDF vectorization, PCA, KMeans clustering, t-SNE visualisation, LDA modelling, and Bokeh visualisation. The grouping and topic modelling of the dataset of BBC stories were highly accurate thanks to our workflow. The results could be interactively seen with the help of the Bokeh visualisation tool, and the large dimensionality of the dataset could be reduced to two dimensions with the aid of the t-SNE technique.

Our primary contribution consisted in fusing these several algorithms to produce a thorough pipeline for text categorisation and visualisation. We also gave a thorough explanation of each algorithm and how it fits into the pipeline as a whole. The high dimensionality of the dataset and choosing the best K-value for KMeans clustering were two difficulties we ran into while doing the experiment. We address these difficulties using, respectively, elbow plots and the t-SNE technique.

In the future, we advise looking into other methods and algorithms that could improve the pipeline's precision and effectiveness. For instance, applying deep learning models like transformers or BERT may increase topic modelling's accuracy. Additionally, experimenting with different visualisation methods and tools could result in data visualisations that are more potent.

Conclusion:

In our study, text is categorised using a combination of conventional machine learning algorithms (TF-IDF vectorization, PCA, KMeans clustering, LDA modelling). Although our method is not as cutting-edge as some of the more current approaches in the literature, it nevertheless offers a straightforward and efficient way to categorise texts.

Individual contributions of each members :

1. Ajay : Conducted data collection and preprocessing, including web scraping, data cleaning, and text normalization. Also, implemented the TF-IDF vectorization and PCA dimensionality reduction methods.

2. Bhavya : Implemented the KMeans clustering algorithm and conducted the elbow method to determine the optimal number of clusters. Also, performed cluster analysis and interpretation.

3. Dinesh : Implemented the t-SNE visualisation technique and created visualisations of the clustering results. Also, implemented LDA for topic modelling and interpretation of the topics.

4. Hemanth: Performed hyper-parameter tuning and optimisation of the clustering and topic modelling methods.

5. Sravya: Developed the Bokeh visualisation dashboard for interactive exploration of the clustering and topic modelling results. Also, contributed to the project report and presentation.

References:

1. A study conducted by Sebastiani (2002) compares different machine learning algorithms for text categorisation. The study concluded that Support Vector Machines (SVM) and k-Nearest Neighbor (kNN) algorithms perform better than other algorithms in terms of accuracy.

2. In 2009, Blei et al. proposed Latent Dirichlet Allocation (LDA), a probabilistic model for topic modelling, which has been widely used in text categorisation.

3. "A Comparison of Document Clustering Techniques" by Eric B. Baum, et al. This paper presents a comparative study of different clustering algorithms for document classification. The authors experiment with several algorithms such as k-means, hierarchical clustering, and expectation-maximization, and evaluate their performance on the Reuters-21578 dataset. The results suggest that k-means is the most effective algorithm for clustering documents.

4. "Efficient and Effective Dimensionality Reduction for Large-Scale and Streaming Data" by Xiang Zhang and Yann LeCun. This paper proposes a new method for dimensionality reduction called "Randomised PCA" that is suitable for large-scale and streaming data. The authors show that their method is computationally efficient and achieves results comparable to traditional PCA methods.

5. "Topic Models" by David M. Blei, et al. This paper introduces the concept of "topic models" as a way to automatically discover the hidden topics in a large corpus of text. The authors propose a probabilistic model called "Latent Dirichlet Allocation" (LDA) that represents documents as mixtures of topics and topics as mixtures of words. LDA has become a popular tool for text classification and information retrieval.

6. "Visualising Data using t-SNE" by Laurens van der Maaten and Geoffrey Hinton. This paper introduces t-SNE, a technique for visualising high-dimensional data in two or three dimensions. The authors demonstrate the effectiveness of t-SNE on several datasets, including text data, and show that it outperforms other techniques such as PCA and LLE for visualising complex structures in the data.