News Article Classification

Group: Classifiers

Yesha Ajudia - AU1841078

Kartavi Baxi - AU1841079

Harsh Kakasaniya - AU1841085

Vimarsh Soni - AU1841121

Introduction

Text classification is the problem of automatically assigning zero, one or more of a predefined set of labels to a given segment of text. The labels are to be chosen to reflect the “meaning” or “context” of the text. Provided Text need to be classified based on various keywords present in the content and their frequency and occurrence defines the category to which the provided text would belong.

Problem Statement

Increase in the digitization has lead to various portals containing information from news, blogs, articles and ebooks. For a frequent user of such digital platforms, it would be really helpful to get desirable information without the need to scroll to large lists and tons of sources. When it comes to news articles, users prefer to read articles/news based on categories that matches to his/her interests. And hence applications and website uses various ways to sort the articles based on different categories which makes it comfortable for users to find them. [1]

Our project focuses on the classification of publicly available BBC news article among 5 selected categories namely: Business, Entertainment, Politics, Sports and Tech, which can be helpful for an e-newspaper reader.

Existing Body of Work

Clean Data:

After text Tokenization, Stop-words Removal and Word Lemmatization

Balanced Dataset:

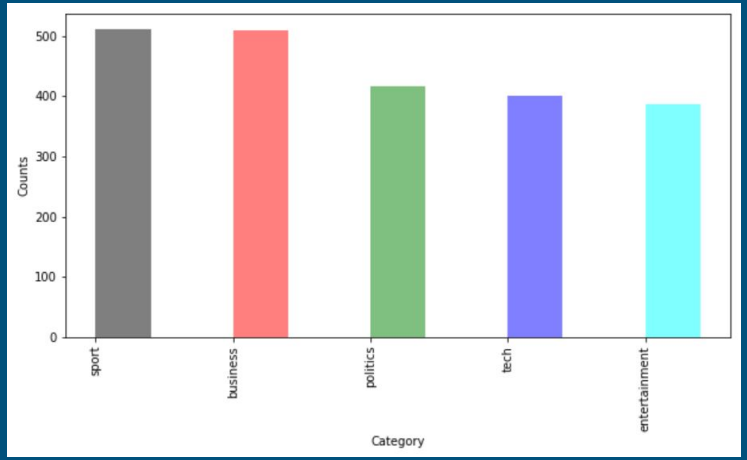
As shown in the figure, the classes are balanced

Unigrams and Bigrams Frequency Distribution:

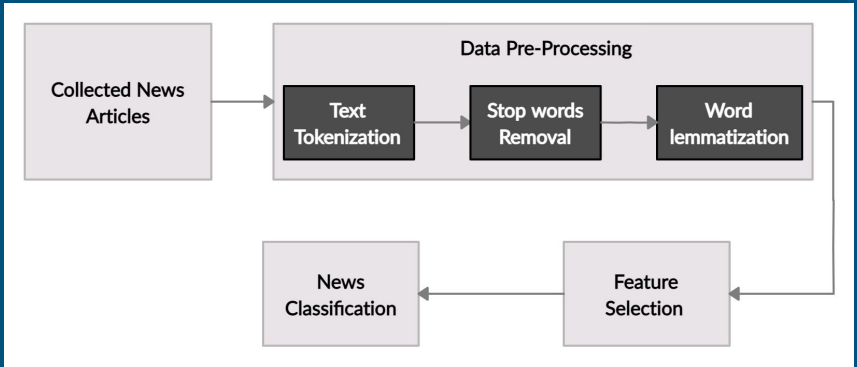
Observe the frequency of related words and its correlation with the category

Train-Test Split:

Dataset splitted into 80% train data and 20% test data



Approach



[2]

News articles were collected from BBC news in the form of dataset consisting of 2225 articles categorized in 5 classes.

Further data cleaning was performed by tokenizing the text, removing stopwords, and lemmatizing them.

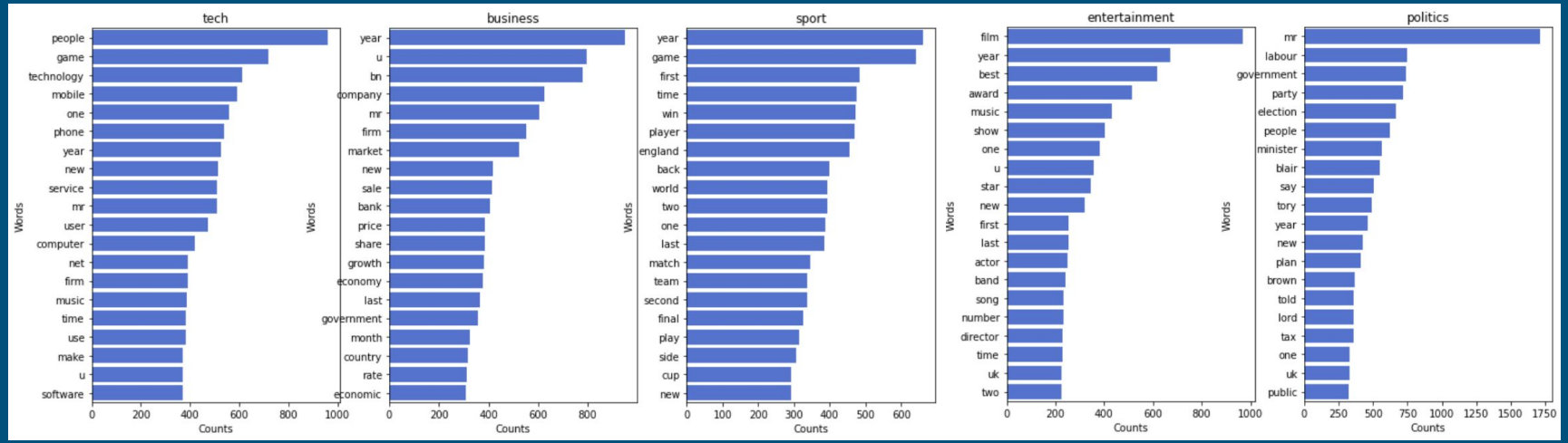
Feature extraction will be performed on the cleaned data and classification algorithm would be applied to get the required results.

Initial Results

From the results obtained, it can be observed that words or unigrams occurring in the content are highly correlated to the category.

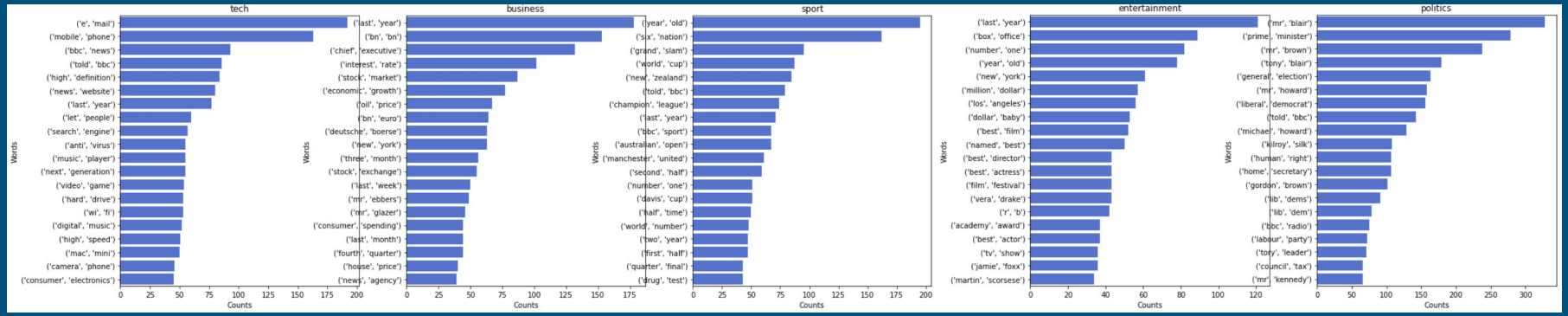
Similarly, for bi-grams, it can be observed that the bi-grams or the pair of words occurring in the content are also highly correlated to the category. Hence, the classification can be done based on the frequency of the words.

Initial Results (Unigram)



From the figures, it can be observed that words or unigrams occurring in the content are highly correlated to the category.

Initial Results (Bigram)



Similarly, from the above figure, it can be observed that the bi-grams or the pair of words occurring in the content are also highly correlated to the category. Hence, the classification can be done based on the frequency of the words.

Role of each group member

Yesha Ajudia

Prepare the CSV from the dataset of text files.

Explore classification algorithms.

Explore feature extraction techniques.

Kartavi Baxi

Harsh Kakasaniya

Explore python libraries required for text classification.

Split train and test data.

Encode the labels to numeric form.

Vimarsh Soni

Find the dataset.

Do the literature review.

Code for data cleaning.

Determine the flow of the project.

Future Work

Extracting features from the cleaned text, using Bag-of-Word and TF-IDF.

Implement the classification algorithms based on features extracted, compare the results and find the best fit model to the data.

References

1. Krishnamoorthy, Arjun, et al. "News Article Classification with Clustering using Semi-Supervised Learning." 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE, 2018.
2. News Classification Based On Their Headlines: A Review : Rana, Mazhar Iqbal, Shehzad Khalid, and Muhammad Usman Akbar. "News classification based on their headlines: A review." 17th IEEE International Multi Topic Conference 2014. IEEE, 2014.

Thank You