School of Engineering and Applied Science (SEAS), Ahmedabad University

BTech(ICT) Semester VI: Machine Learning (CSE 523)
Mid Semester Project Report

News Article Classification

Yesha Ajudia
AU1841078

Kartavi Baxi
AU1841079

Harsh Kakasaniya
AU1841085

Vimarsh Soni
AU1841121

***Abstract*—This project is based around a supervised Machine Learning Text Classification model coded in Python which would be able to predict the category of a given news article from the predefined set of categories. It would use a labelled data-set helping the algorithm to learn the patterns and correlations in the data. Data Cleaning would be used to ensure no distortions to the model. Feature extraction techniques based on the frequency of uni-gram and bi-gram would be used to analyze the category of the given news article. This would help us to analyze patterns in the data and gain valuable insights from it.**

***Index Terms*—Text Classification; Category Identification; News Article; Data Pre-processing; Supervised Machine Learning**

I. Introduction

Increase in the digitization and the number of phone and similar users has lead to various portals containing information from news, blogs, articles and e-books. 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]

This paper focuses on the classification of publicly available BBC news articles dataset among 5 predefined categories namely: Business, Entertainment, Politics, Sports and Tech, which can be helpful for an e-newspaper reader. The dataset consists of content of the article and its category, which can be visualised as a key-value pair with article being the key and category being the value. The classification model characterizes each article based on the features extracted. Frequency distribution of words in the content is used in the feature extraction techniques. Labelled data-set is used to help the algorithm to learn the patterns for those occurrences in the data. Every provided article will be analysed and its words list frequency will be evaluated to classify it in one of the 5 pre-defined categories.

II. Literature survey

Text pre-processing is a vital stage in text classification particularly and text mining generally. It can be done by text documents collection, tokenization, stop-words removal, and stemming [4]. The frequency distribution of every uni-gram and bi-gram gives the idea about the frequent words appearing in all categories. N-grams are a viable alternative to words as indexing terms in information retrieval. N-grams provide higher accuracy than a strawman system using raw words as indexing terms [5]. Data splitting is dividing the dataset into two subsets – one subset is used for training while the other subset is left out and the performance of the final model is evaluated on it. Data splitting is performed to train the model on a specific data set and then test it on unseen data to get best understanding of accuracy [6]. All the classification algorithms has its own pros and cons and a good performance on classification demands the right choice of classifier for the right problem [10]. Bag of Words just creates a set of vectors containing the count of word occurrences in the document, while the TF-IDF model contains information on the more important words and the less important ones as well. Bag of Words vectors are easy to interpret. However, TF-IDF usually performs better in machine learning models [11].

III. Implementation

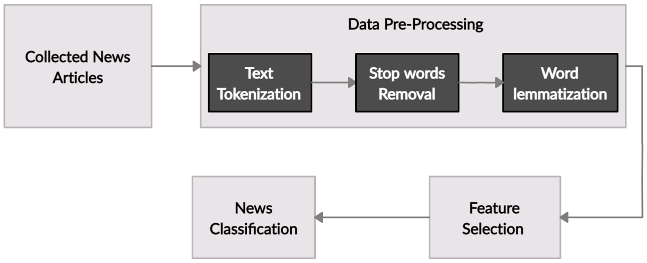
Along with having a literature survey on the machine learning classification algorithms and the feature extraction techniques, so far, we have performed the following steps: [2]

1) DATA CLEANING:
   This step is perfromed to remove the noise from the content. The content in the article might consist of some short form notations like won't, can't, etc, which needs to be converted to their full forms in order to help in stop-words removal. Any character other than alphabets (digits, special symbols, etc) are removed and the whole text is converted to lower case in order to avoid case-sensitivity. The cleaned content is then stored in a separate column. Next step removes the stop-words like 'i', 'me', 'my', 'myself', etc from the content, using the stop-words present in the nltk library and lemmatizes the text. Lemmatization means extracting the root word from the given word. It is required as a single word may be used in the content in different forms. Hence, going to the root word becomes important.
   Some words like 'would', 'could', 'also', etc were not removed by using the in-built stop-words list. Hence, such words were removed manually after observation.

2) CHECK BALANCE:
It becomes important to verify the balance in the classes while working on classification problems, as an imbalance in them would lead to mis-classification [3].

3) UNI-GRAMS and BI-GRAMS:
A uni-gram is defined as a word that individually occurs in the content. A bi-gram is defined as a pair of two words taken at a time that occur in the content. Bi-grams are considered for words like mobile phone, last year, e-mail, New York, etc. that generally occur together in any content. Once the data has been cleaned after all the pre-processing steps, the content of the article now have words which are highly correlated to the category. Hence, to observe the words and their frequency, we have plotted top 20 uni-grams and bi-grams occurring in the cleaned content. Fig. 3 shows the top 20 uni-grams and Fig. 4 shows the top 20 bi-grams occurring in the articles of the 5 categories.

4) TRAIN TEST SPLIT:
Next we split the dataset into 80% train data and 20% test data, after encoding the labels to numeric form (as machines do not understand non-numeric text).



The figure above, Fig. 1 shows the intended flow of the project.

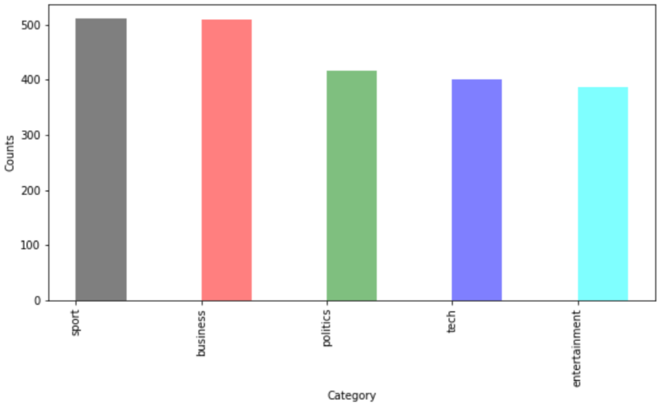


Fig. 1. Category-wise Article Distribution

IV. Results

The results generated till now include: the cleaned text after performing the data cleaning process (which holds a major part when working on any text classification problem), a balanced dataset, the frequency distribution plots of uni-grams and bi-grams, train and test data separated. Insights obtained from literature survey include how to perform the step-wise data cleaning process, why is a balanced dataset important while working on classification problems, why should we limit our model to uni-grams and bi-grams without going to a higher n-gram model, why should the labels be encoded to numeric form, about the feature extraction techniques, and classification algorithms.

Labelled dataset of nearly 1800 samples (80% of the whole dataset) will be used to train the model to predict the category of a given article. Every new article provided to algorithm will be classified into one of these five categories based on the features extracted using feature extraction techniques like Bag-Of-Word and TF-IDF.

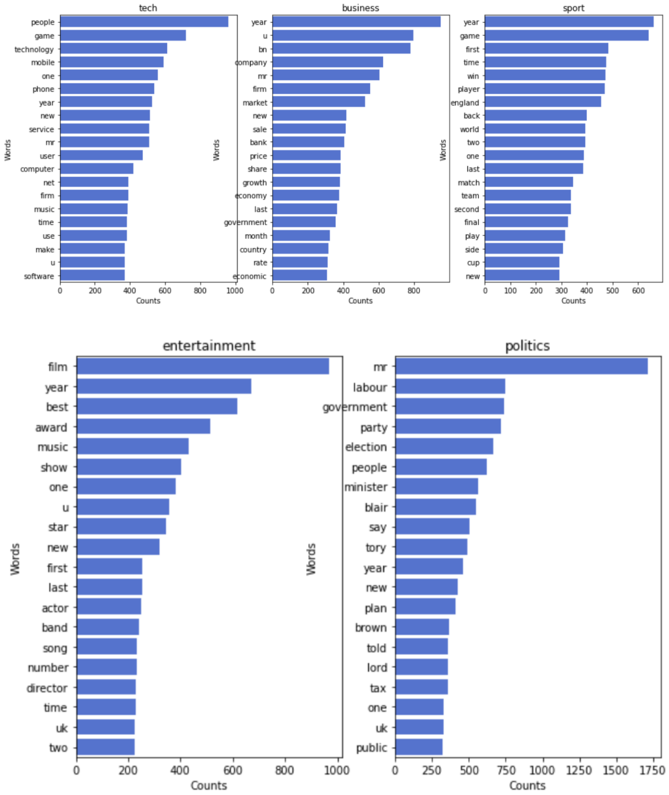


Fig. 2. Uni-gram Frequency

V. Conclusions

This paper explains the detailed work done for news article classification including the steps performed till now which are, data collection and pre-processing, cleaning of data, observing frequency distribution of uni-grams and bi-grams, encoding, and splitting of train and test data. It also discusses the removal of unused words from the data such as conjunctions, pronoun and prepositions due to their negligible contribution in the classification of article in 5 categories. In future, we would be working on the main classification algorithm and feature extraction techniques, to be implemented on the training data to get the model ready to classify the provided article in one

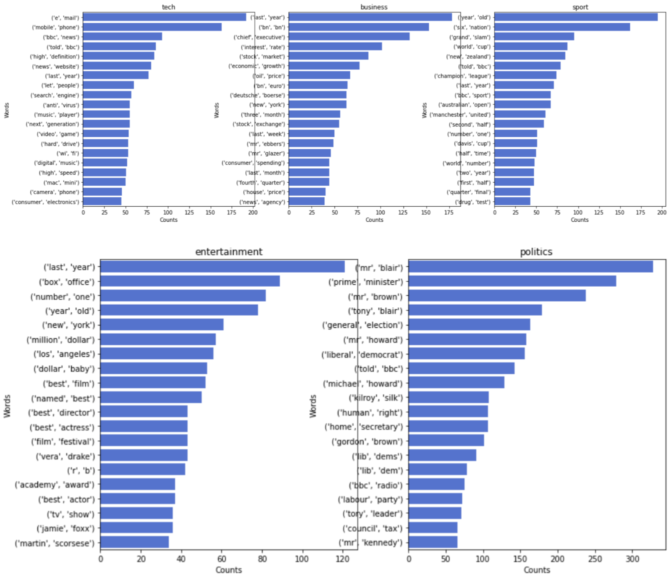


Fig. 3. Bi-gram Frequency

of the five (Tech, Sports, Business, Politics and Entertainment) categories.

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] Kaur, Gurmeet, and Karan Bajaj. ”News classification and its techniques: a review.” IOSR Journal of Computer Engineering (IOSR-JCE) 18.1 (2016): 22-26.

[3] Tripathi, Himanshu. “What Is Balanced And Imbalanced Dataset? - Analytics Vidhya.” Medium, 24 Sept. 2019, medium.com/analytics-vidhya/what-is-balance-and-imbalance-dataset-89e8d7f46bc5.

[4] Kadhim, Ammar Ismael. ”An Evaluation of Preprocessing Techniques for Text Classification.” International Journal of Computer Science and Information Security 16.6 (2018).

[5] McNamee, P., Mayfield, J. Character N-Gram Tokenization for European Language Text Retrieval. Information Retrieval 7, 73–97 (2004).

[6] Reitermanova, Zuzana. ”Data splitting.” WDS. Vol. 10. 2010.

[7] Ikonomakis, M., Sotiris Kotsiantis, and V. Tampakas. ”Text classification using machine learning techniques.” WSEAS transactions on computers 4.8 (2005): 966-974.

[8] Agarwal, Basant, and Namita Mittal. ”Text classification using machine learning methods-a survey.” Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012. Springer, New Delhi, 2014.

[9] 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.

[10] V. K. Vijayan, K. R. Bindu and L. Parameswaran, ”A comprehensive study of text classification algorithms,” 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Udupi, India, 2017, pp. 1109-1113, doi: 10.1109/ICACCI.2017.8125990.

[11] Text, Q. and Huilgol, P., 2021. BoW Model and TF-IDF For Creating Feature From Text. [online] Analytics Vidhya. Available at: ¡https://www.analyticsvidhya.com/blog/2020/02/quick-introduction-bag-of-words-bow-tf-idf/¿ [Accessed 17 March 2021].