2020 Fall Data Mining (MGSC-5126-10)

Karthik Selvaraj (20192394) & Gokul Nagarajan (20193832)

NEWS HEADLINE CLASSIFICATION

ABSTRACT:

The ability to automatically identify and model documents into a defined set of categories is highly beneficial in real-world scenarios. This categorization of news articles presents a major challenge for user-driven news classification applications that are involved in assessing their users' preferences and thereby delivering the most appropriate information. We plan to simplify the classification of news articles by using machine learning & Natural Language Processing (NLP) techniques. We discuss three methods of classification: Naïve Bayes, Support Vector Machine (SVM) & SoftMax Regression, and test the ability of each classifier to pick the appropriate category provided the title of an article and a brief overview of the article. Our findings reveal that Support Vector Machines as the best classifier among the three other classifiers we tested.

Index Terms: News, Articles, NLP, SVM, Naïve Bayes, Neural Networks, Machine Learning, SoftMax, Classification, TF-IDF

INTRODUCTION:

There are numerous news sources, online news portals that let out every minute of the daily operation. News reports are relevant and flow everywhere, whether it's print media or electronic media. Therefore, it is important to have an effective method of segregating news into various categories (Rana et. al., 2014). In contrast to those in news headlines, prospects for misclassification in the classification of news stories with broad details are more common. Numerous researchers are proposing approaches to classifying news headlines that classify each news headline into its pre-defined class (Deb et. al., 2020). The news headlines will be used to train the model and the computer will be able to very easily and correctly predict the category of the news item later. For all news channels and apps, this will be beneficial as it will give them an easy and speedy way to do their work.

RESEARCH OBJECTIVES:

The aim of this research project is to identify and predict news headlines, evaluate them and compare the findings using different machine learning models. We will test and consider the ability of classifiers to pick and assess the article type, provided the title and brief description of the article.

Collection of news, preprocessing of collected news, selection of features, different classification techniques for classifying news and assessing performance measures for different classification techniques are the various steps involved in news classification.

LITERATURE REVIEW:

The key task of news classification is to automatically classify the news documents into their predefined groups based on their content. Many methods of machine learning for classifying news have been developed. In the field of text mining, classification is a difficult task since it involves preprocessing steps to prepare textual data in a structured form that is initially available in an unstructured form (Rana et. al., 2014).

The feature selections in this research are done by using TFIDF. Term Frequency-Inverse Document Frequency (TF-IDF) (Wu et. al., 2008) is a very common algorithm which is used to transform text into a meaningful representation of numbers.TF-IDF can be used for stop-words filtering in various subject fields including text summarization and classification.

The next most significant stage after the collection of features is the classification of the news headlines in order to allocate them to their respective classes (Deb et al., 2020). Naïve Bayes, Support Vector Machine and Neural Network with SoftMax Layer would be the classification methods for news headlines in our project.

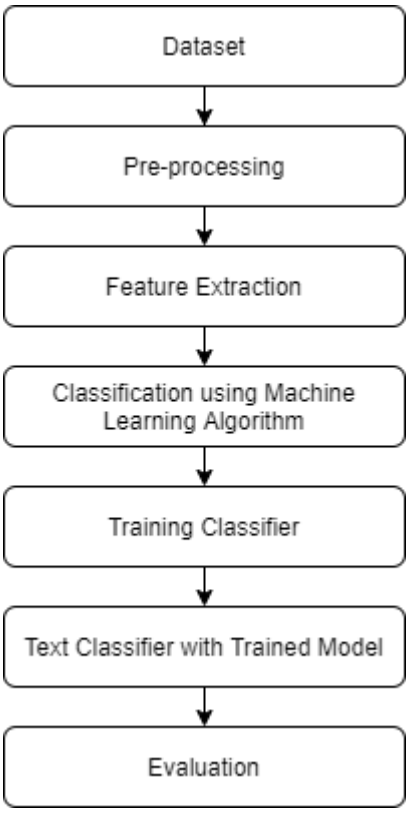
A new feature scaling approach that uses the Naïve Bayes classifier was introduced by Young and Jeong. On a news community dataset, the feature scaling approach was evaluated and outperformed other common rating schemes, such as Information Gain, while Naïve Bayes was noted as a suitable classifier for news posts (Youn & Jeong, 2009). An efficient text categorization algorithm based on the SVM algorithm used in this paper was developed by Wang et al (Wang et al., 2006). They find their algorithm to outperform other classifiers, such as the decision-tree algorithm and the K-nearest neighbour algorithm, by using a news article corpus like ours.

The potential of the Word Frequency-Inverse Document Frequency (TF-IDF) algorithm to be used in text classification for Bahasa Indonesia newspapers was evaluated by Hakim et al (Hakim et al., 2014). However, their methodology did not concentrate on any methods for machine learning, only on the TF-IDF algorithm.

Ruiz and Srinivasan developed one of the first architectures for Neural Networks. Using around 2,350 texts, they illustrated Neural Networks' capacity to categorize text correctly. Using a

modified SoftMax Regression algorithm, Do and Ng explored text classification (Do & Ng, 2005).

PIPELINE DIAGRAM:



DATASET & PRE-PROCESSING:

News Collection

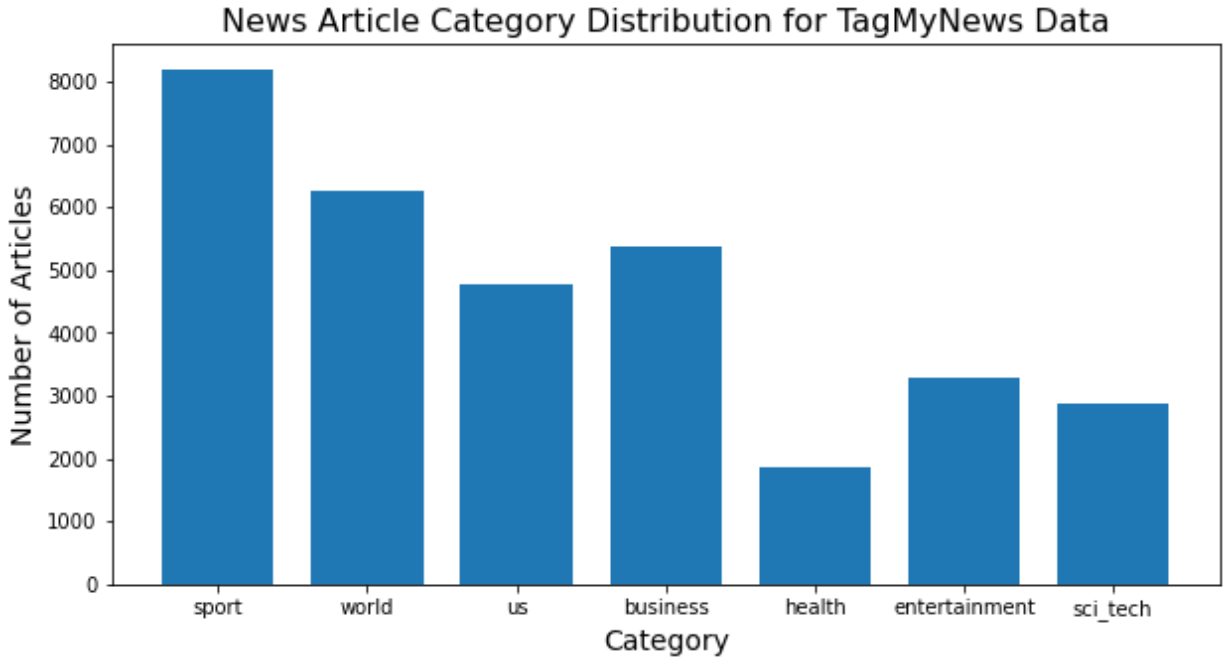
In this work, we will be using a dataset from TagMyNews which is a collection of datasets of short text fragments for the evaluation of our topic-based text classifier. Our dataset contains 32K English news extracted from RSS feeds of popular newspaper websites (nyt.com, usatoday.com, reuters.com). Categories are: Sport, Business, U.S., Health, Science & Technology, World and Entertainment (TagMyNews Datasets, 2017)

Each news in the file has the following structure:

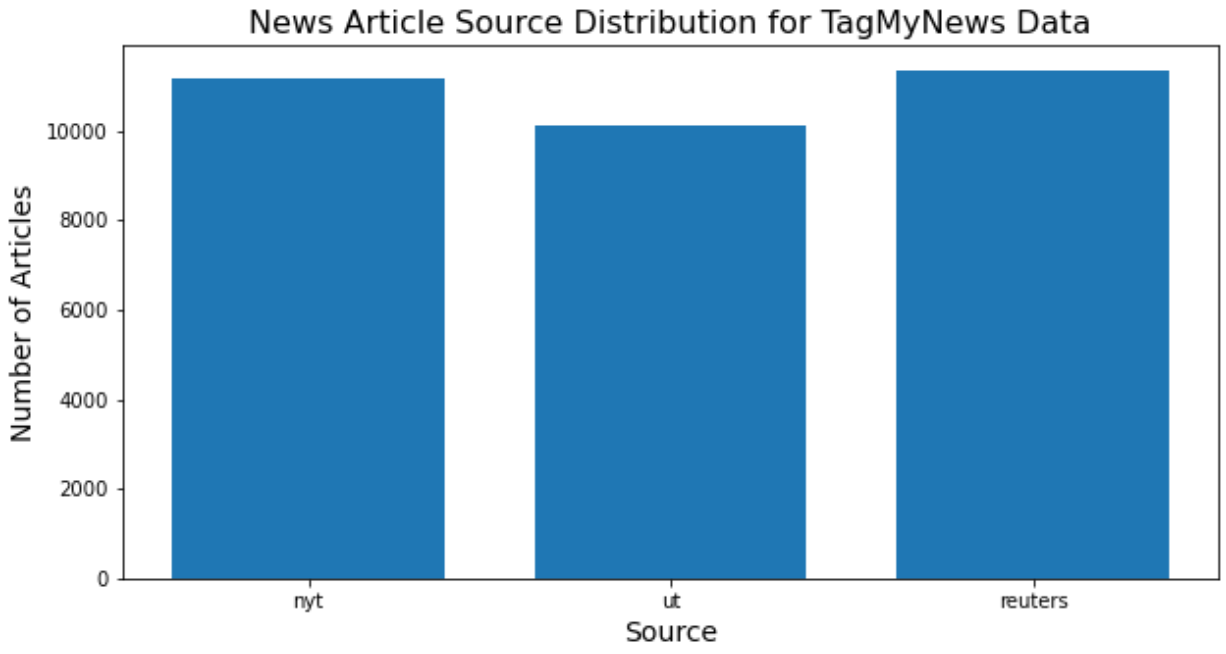
Variable	Description
Title	Title of the Article
Description	A short description about the article
Link	URL to access the article
ID	Article ID
Date	Publication Date
Source	New York Times/ USA Today/Reuters
Category	Sport, Business, U.S., Health, Science & Technology, World and Entertainment.

Exploratory Data Analysis

Initial exploratory data analysis revealed the number of articles in each category which is shown as follows. From the analysis we found that 'sport' category has the highest number of articles.



An interesting analysis revealed that from the dataset, Reuters news source has the highest number of articles followed by New York Times and USA Today.



News Pre-Processing

Pre-processing is performed after news gathering, since this material originates from several sources and its cleaning is important so that it can be free from corrupted archives. 'news' file with the data is opened and the text is read. Using the split function, each news item is separated and stored. Various categories of the news articles are recognized and their respective count is initialized from 0. For each news item, the data is written and stored with their respective filename in their folder category and closed after the written process.

Directory list is defined which is used by the glob module to retrieve files matching the pattern. Article Title with their Description along with their category has been fetched and store in the file for our further processing. Categories are flagged from 0 to 6 where 0: sport, 1:world, 2: us, 3:business, 4:health, 5:entertainment, 6:sci_tech

Feature Extraction

TFIDF is a statistical measure that assesses how important a word is to a document in a series of documents. The TF-IDF weighting scheme will assign each term, t, a given weight in a document as follows:

$$w_{t,d} = tf_{t,d}\ x\ \log(N/df_t)$$

where N is the number of documents. The weight is allocated by the product of $tf_{t,d}$, the term frequency, and $\log(N / df_t)$, the inverse document frequency.

Count Vectorizer is used to get the vector count from the collection of news_data. It enables the pre-processing of text data prior to generate the vector representation. This functionality makes it a highly flexible feature representation module of a text. Stopwords are defined along with the analyzer and the token_pattern. The word vector is then saved to pickle file which serializes the object before writing it to the file. This serialized object contains all the information necessary to reconstruct the object in another python script.

The word vector obtained from the count vectorizer is then transformed using a TFIDF transformer which transforms a count matrix to a normalized tf or tf-idf representation. TFIDF transformer calculates the inverse document frequencies and it encodes the documents. The transformed TFIDF data is then stored in the pickle file and used as input to the classifiers.

RESEARCH METHODOLOGY:

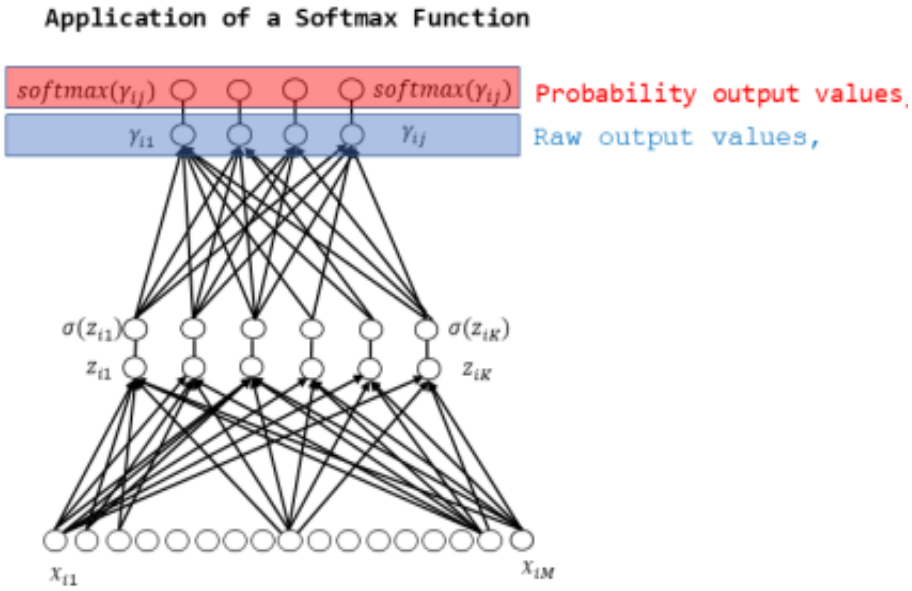
Data has been split into training news articles and testing news articles. 75% of our data (24,453 articles) were designated as the training articles and the remaining 25% (8151 articles) were designated as the testing articles. Thanks to their ability to carry out supervised learning on multi-class datasets, we decided to test Multinomial Naïve Bayes, SVM, and SoftMax Regression.

The Multinomial Naive Bayes applies the Naive Bayes algorithm for multinomially distributed data and the classifier is appropriate for discrete attribute classification (e.g., word counts for text classification). Typically, the multinomial distribution includes integer function counts and fraction counts which are used in operation, such as TF-IDF.

Linear SVC (Support Vector Classifier) matches the information supplied and returns a

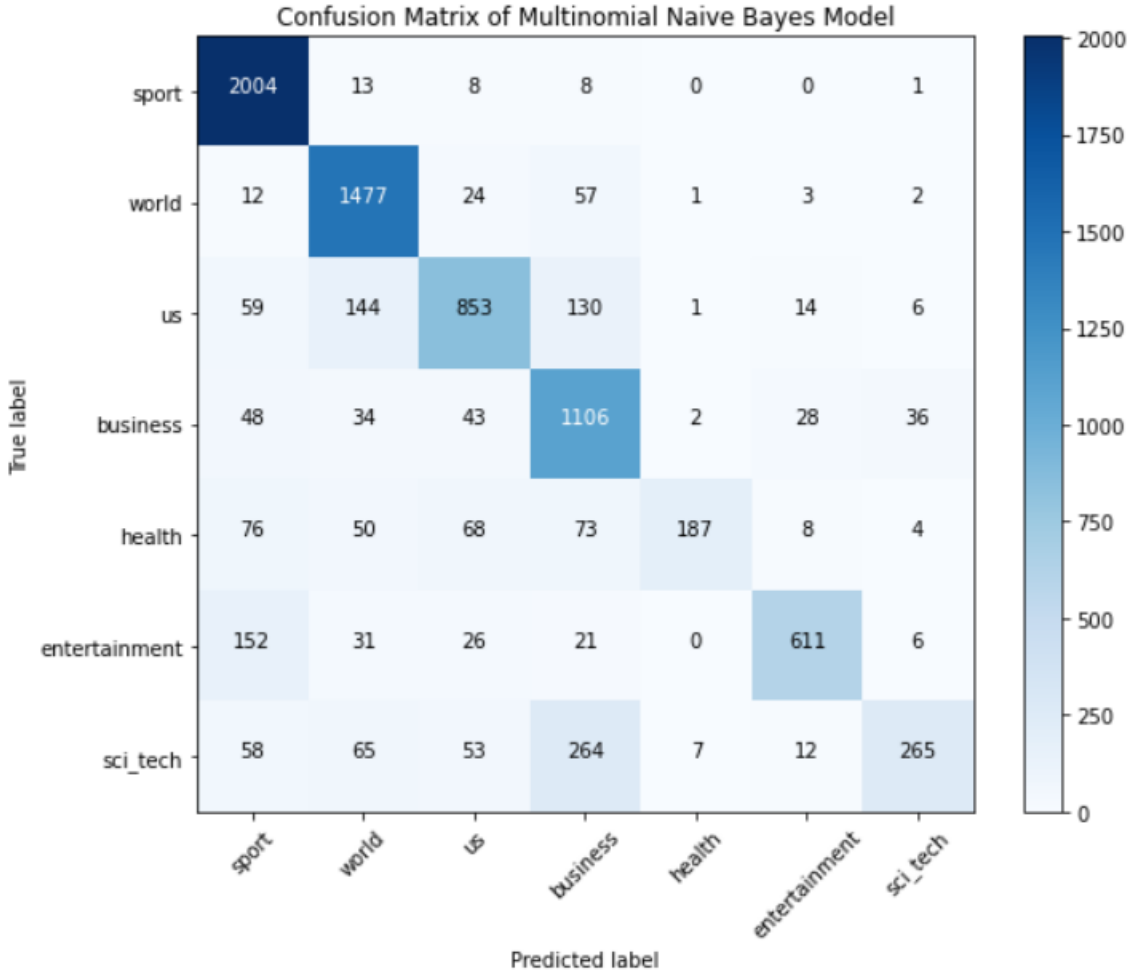
hyperplane "best fit" that divides or categorizes the data. Some features are fed to the classifier after receiving the hyperplane in order to achieve the predicted class.

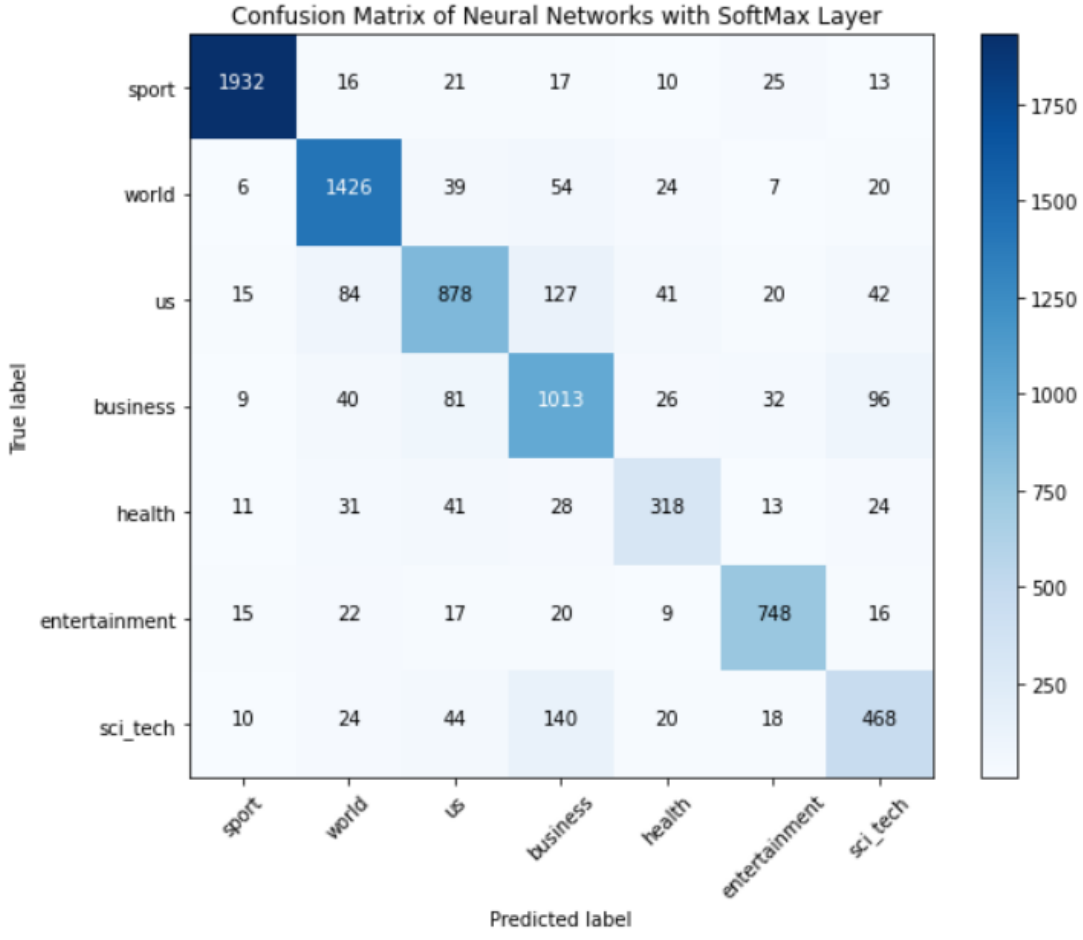
The SoftMax feature is commonly used in a neural network-based classifier's final layer. It is a generalization of logistic regression that we can use for the classification of multiple classes. In contrast with 7-binary classifiers, we preferred SoftMax Regression, also known as the Multinomial Logistic Regression, since our seven classes are mutually exclusive (i.e. a news article will be a part of at most one category).

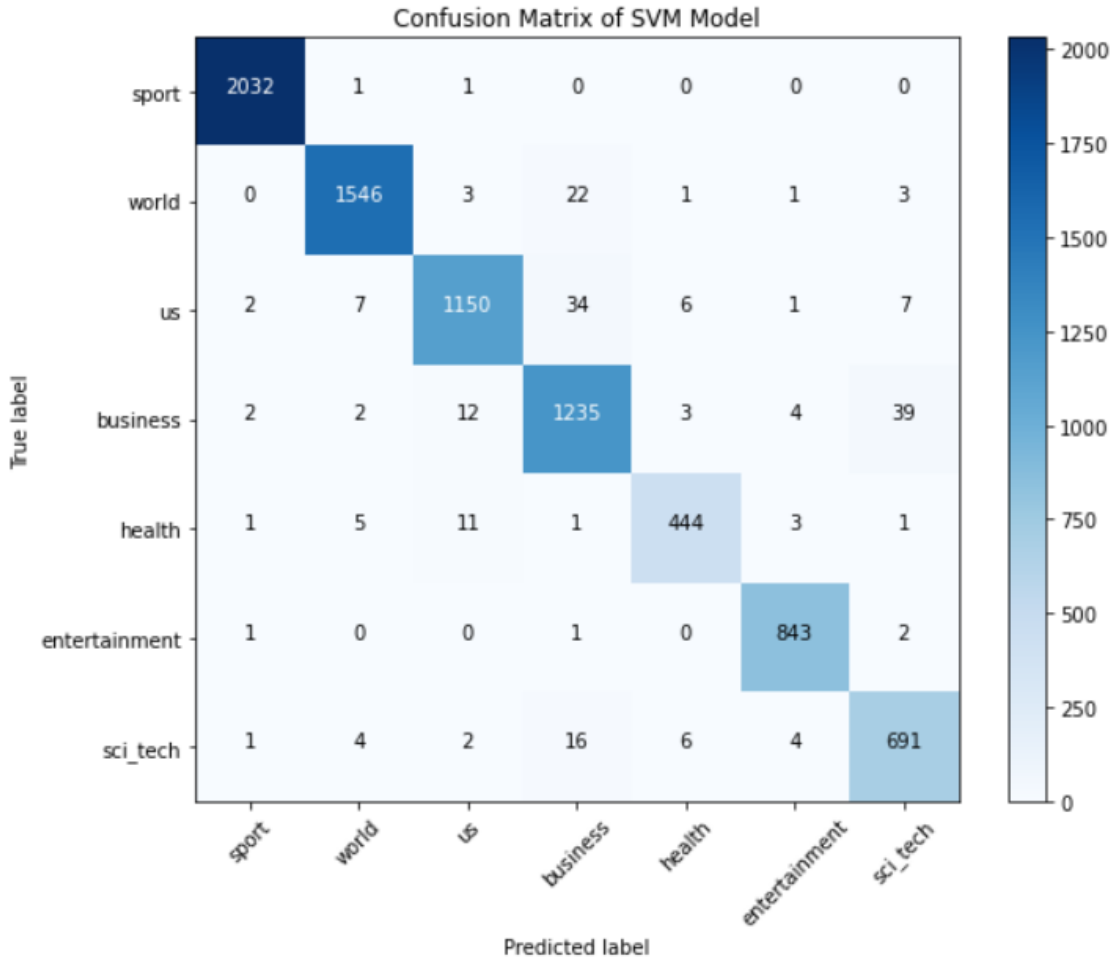


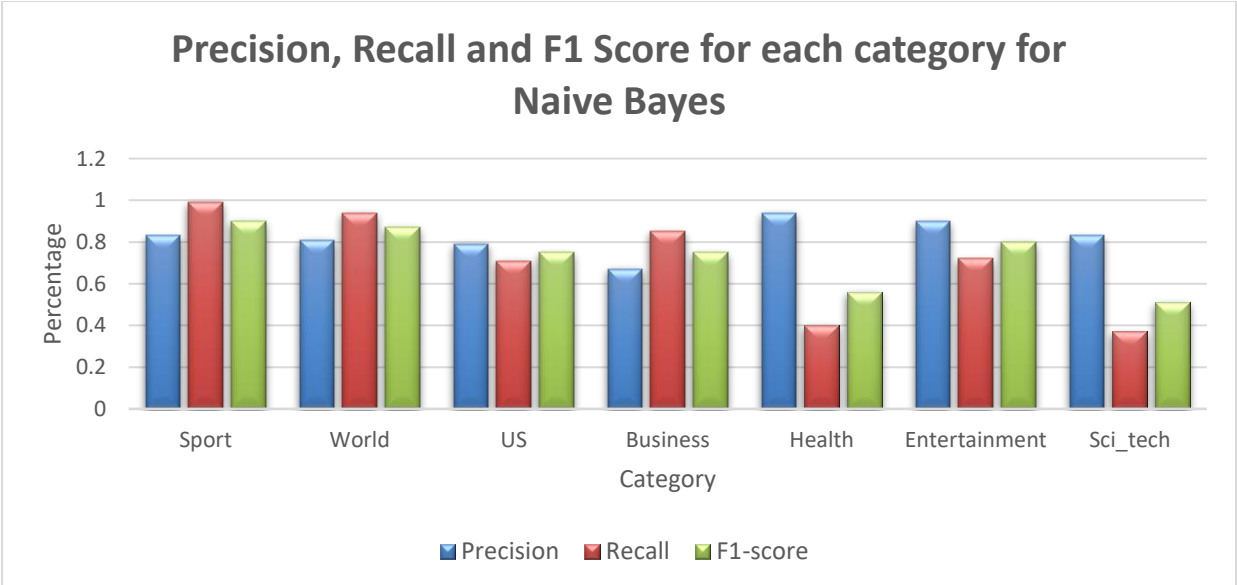
RESULTS & CONCLUSION:

We used the three metrics derived from the confusion matrix to further test the utility of each classifier on individual classes: accuracy, recall, and F1-score. Precision is the class agreement of the data labels with the classifier's positive labels, while recall is the classifier's effectiveness in detecting positive labels. The harmonic mean of precision and recall is the F1-score. Nevertheless, we have shown the ability of three different classifiers to automatically classify news articles into their subject category.

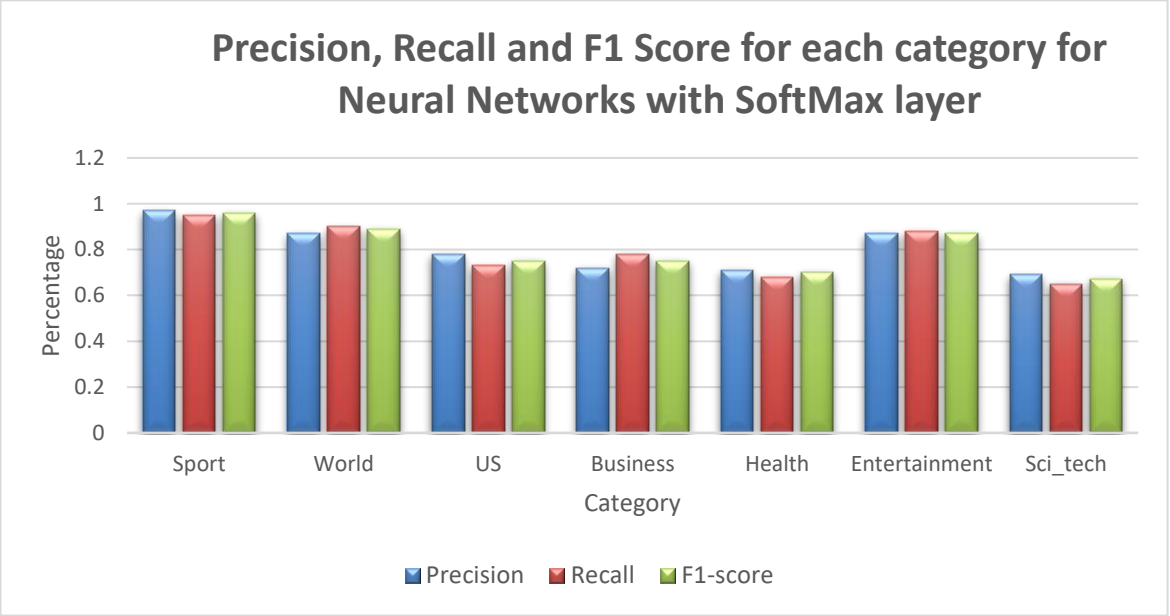

Confusion Matrix of Multinomial Naive Bayes Model
sport 2004 13 8 8 0 0 1
world 12 1477 24 57 1 3 2
us 59 144 853 130 1 14 6
business 48 34 43 1106 2 28 36
health 76 50 68 73 187 8 4
entertainment 152 31 26 21 0 611 6
sci_tech 58 65 53 264 7 12 265
sport world us business health entertainment sci_tech
True label
Predicted label
2000 1750 1500 1250 1000 750 500 250 0


Confusion Matrix of Neural Networks with SoftMax Layer
sport 1932 16 21 17 10 25 13
world 6 1426 39 54 24 7 20
us 15 84 878 127 41 20 42
business 9 40 81 1013 26 32 96
health 11 31 41 28 318 13 24
entertainment 15 22 17 20 9 748 16
sci_tech 10 24 44 140 20 18 468
sport world us business health entertainment sci_tech
True label
Predicted label
1750 1500 1250 1000 750 500 250

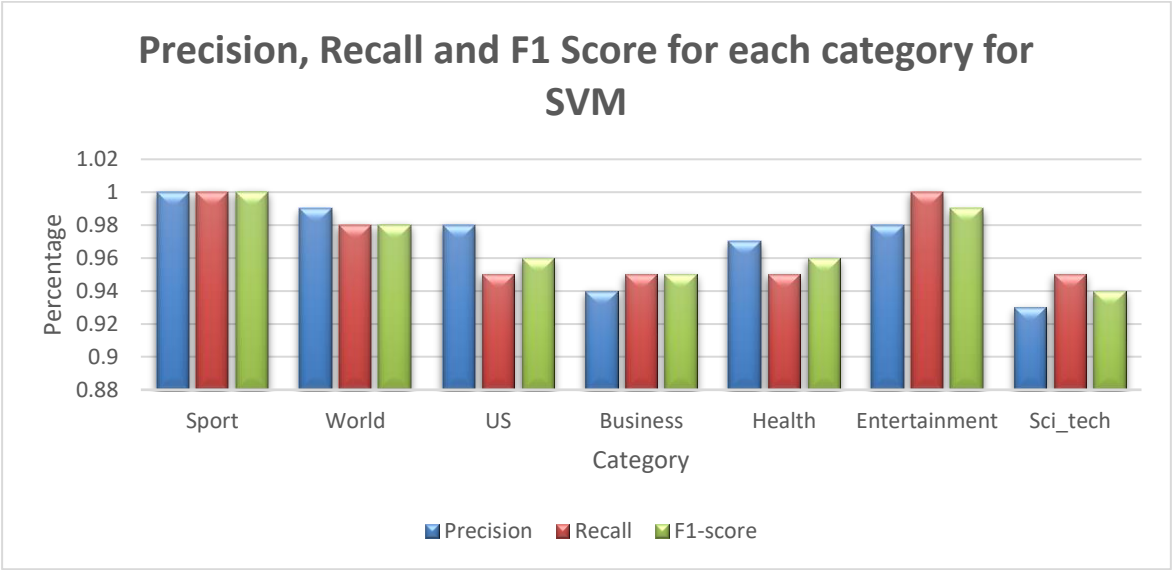




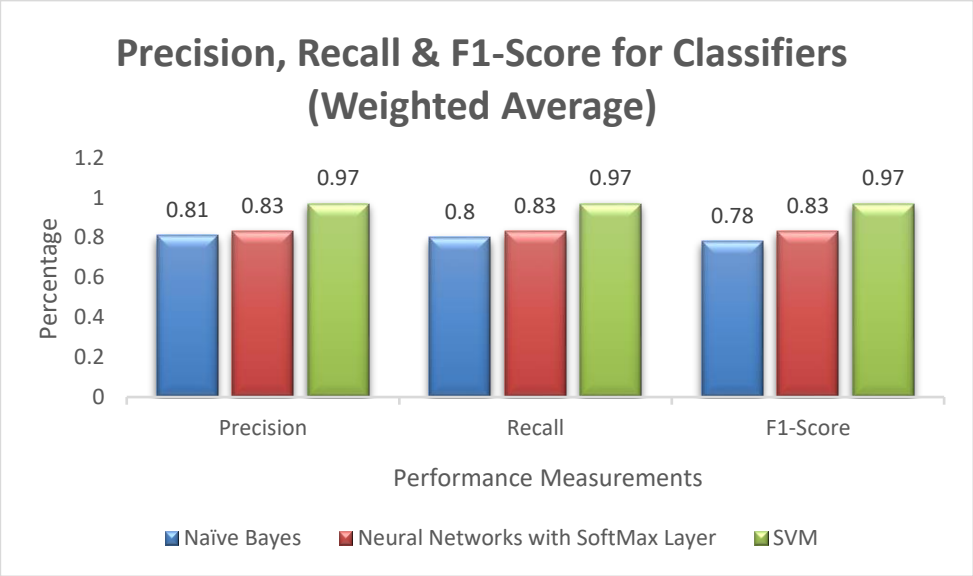
The above graph shows the precision, recall & F1-score for each category for Multinomial Naïve Bayes classifier. Precision for the health & entertainment category is higher compared to other categories whereas recall and F1-score is higher for sport & world categories.



The above graph shows the precision, recall & F1-score for each category for Neural Networks with SoftMax Layer. Precision for the sport & entertainment category is higher compared to other categories whereas recall and F1-score is higher for sport & world categories.



The above graph shows the precision, recall & F1-score for each category for Support Vector Machines. Precision for the sport & world category is higher compared to other categories whereas recall and F1-score is higher for sport & entertainment categories.



From the above graph, it can be deduced that SVM has the highest precision, recall & f1-score among all the three classifiers followed by Neural Networks with SoftMax Layer & Naïve Bayes classifier. SVM also has the highest accuracy when compared among all the other classifiers. Hence, SVM classifier was found to be highly suitable for classifying news article into their subject category in our case.

REFERENCES:

1. Deb, N., Jha, V., Panjiyar, A. and Gupta, R. (2020). A Comparative Analysis of News Categorization Using Machine Learning Approaches. [online] Available at: http://www.ijstr.org/final-print/jan2020/A-Comparative-Analysis-Of-News-Categorization-Using-Machine-Learning-Approaches.pdf [Accessed 20 Oct. 2020].

2. Do, C.B. and Ng, A.Y., 2005. Transfer learning for text classification. Advances in neural information processing systems, 18, pp.299-306.

3. Hakim, A.A., Erwin, A., Eng, K.I., Galinium, M. and Muliady, W., 2014, October. Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach. In 2014 6th International Conference on Information Technology and Electrical Engineering (ICITEE) (pp. 1-4). IEEE.

4. Rana, M.I., Khalid, S. and Akbar, M.U., 2014, December. News classification based on their headlines: A review. In 17th IEEE International Multi Topic Conference 2014 (pp. 211-216). IEEE.

5. Ruiz, M.E. and Srinivasan, P., 1998, October. Automatic text categorization using neural networks. In Proceedings of the 8th ASIS SIG/CR Workshop on Classification

Research (pp. 59-72).

6. TagMyNews Datasets. Web.Archive.Org, 26 Apr. 2017, web.archive.org/web/20170426015149/acube.di.unipi.it:80/tmn-dataset/. Accessed 8th Oct. 2020.
7. Wang, Z.Q., Sun, X., Zhang, D.X. and Li, X., 2006, August. An optimal SVM-based text classification algorithm. In 2006 International Conference on Machine Learning and Cybernetics (pp. 1378-1381). IEEE.
8. Wu, H.C., Luk, R.W.P., Wong, K.F. and Kwok, K.L., 2008. Interpreting TF-IDF term weights as making relevance decisions. ACM Transactions on Information Systems (TOIS), 26(3), pp.1-37.
9. Youn, E. and Jeong, M.K., 2009. Class dependent feature scaling method using naive Bayes classifier for text datamining. Pattern Recognition Letters, 30(5), pp.477-485.