International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 06 | June 2021 www.irjet.net p-ISSN: 2395-0072

Real Time News Headlines Classification Using Machine Learning

Manvendra Singh Chhajerh[1], Ananya KVS[2], Prof. Merin Meleet[3], Dr. Rajashekara Murthy S[4]

[1,2,3,4]*Department of Information Science and Engineering, RV College of Engineering, Bengaluru*

Abstract - *The online news portals consist of several types of information entering from various sources. In most real-life scenarios, it is greatly desirable to classify this information in an appropriate set of categories and it is important to have an efficient system of segregating news into different groups. Machine Learning is used to enhance and improve the system of classification. Research in the domain of news headlines classification is superficial, and this leads to an opportunity to analyze this topic in greater depth. This paper focuses on real time news classification on the basis of its headlines. A system has been designed to classify each news headline to its pre- defined category. The model is trained such that the machine is be able to predict the category of the news item accurately. A hybrid model based on different algorithms has been created to increase the accuracy of standalone algorithms. The news headline will be fetched in real time and will be passed through this classifier. This whole process will not only lead to a better working model but also show a comparative study of different models for classifying news headline.*

Key Words: Machine Learning, Text Mining, Classification, Supervised learning, Multinomial Naïve Bayes, Support Vector Machine, Logistic Regression

1.INTRODUCTION

The news world is buzzing every moment with news arriving from all sources. There are many online news portals, news channels which let out everyday proceedings each and every minute. Different varieties of news make it to the online portals. It is essential to have an effective system of segregating news into various groups/categories. The use of technology in particular Machine Learning can enhance and improve this system.

This work is aimed at real time news classification on the basis of its headlines. Researchers have worked a lot for carrying out news classification at full text level but work in the domain of news headlines classification exists in a very limited ratio [2]. This leads to a big opportunity to analyze it in greater depth and to construct a better accuracy model. This ranges from tuning of different hyperparameters to using a different algorithm in a different structure for better results. Depending on the availability of headlines, a framework will be designed which classifies all the news headlines to the predefined category. The news headlines will be used for the purpose of training the model and the m/c will predict the category of news item in an active and accurate manner.

For the purpose of this work, datasets from HuffPost, UCI news and a dataset from TagMyNews have been used. The dataset consists of news headlines, short description of news and few additional information such as author and title.

It also contains metadata such as the source of news and date of publication. In this paper, supervised machine learning algorithms namely Multinomial Naïve Bayes (mNB), Logistic Regression (LR), Support Vector Machine (SVM) and Neural Network (NN) has been incorporated and compared in terms of their efficiency to provide accurate results.

2. OVERVIEW OF DIFFERENT TECHNOLOGIES POTENT FOR NEWS CLASSIFICATION

This section consists of the theories and concepts used for implementing the classification of news headlines.

2.1 Machine Learning

ML is a part of artificial intelligence and is based on the theory that systems can learn from data and identify different patterns along with making decisions with nominal human intervention. Machine Learning has proven to be one among the most revolutionary technological innovations in the last decade. Even in this progressively driven corporate world, it is facilitating companies to accelerate digital transformation. Machine learning utilizes algorithms and statistical models to visualize the information and draw conclusions from various patterns that are analysed from the data. It is used to build systems which are capable of learning and acclimatizing themselves without explicit instructions. Machine learning algorithms are the mechanisms of ML since it is the algorithms that turn a data set into a model. ML algorithms are mainly categorized into supervised learning, unsupervised learning, and reinforcement learning. In this work, the algorithms namely, Multinomial Naïve Bayes, Logistic Regression, Support Vector Machines and Neural Networks have been utilised.

2.2 Text Classification

Countless data which does not have a particular structure is pouring in from various channels on the internet such as websites, news portals, survey channels, emails, online reviews etc. and it is difficult to extract useful information from this data because it is not organized in any specific way. Manual filtering by experts requires enormous amount of time and resources to sort the data. Text classification provides a structure and an overall view to this data in an economical and scalable way. With the help of text classification, businesses are able to obtain insights from data which is allowing the automation of business processes. Text classification is known by various terms such as text tagging or text categorization and this is a machine learning technique of categorizing text into organized groups. Text can be automatically analyzed, and they are assigned a set of pre- defined tags/categories based on the content with the help of a text classifier. The first process is to transform text into something that is understandable by a machine which is usually conducted by using a bag of words. Here, the frequency of a word is represented by a vector. After the process of data vectorization, the training data consisting of feature vectors for individual text samples and tag is supplied into the text classifier model. The model is able to make precise predictions with adequate training samples.

2.3 Text Mining

The enormous volume of data generated by businesses every day presents helps businesses to get keen insights on customer's views about a product or a service but there is the challenge of processing all this data. Text mining plays an important role in this area. Businesses are capable of analysing significant and complex data in a quick, simple, and effective manner by virtue of text mining. Text mining helps to reduce manual and repetitive tasks and saves precious time.

Text mining is an artificial intelligence technology that makes use of NLP to process textual data. Unstructured data/text present in databases as well as documents is transformed into normalized and structured data which is suitable to drive. It identifies relevant information within a text that would otherwise remain buried in the mass of textual big data and therefore, it provides qualitative results. Automated text analysis is feasible when Machine Learning and Text Mining is combined. Techniques of text mining include Word frequency, Collocation and Concordance. Word frequency is used to recognize the most recurring terms in a collection of data. Collocation such as bigrams and trigrams improve the granularity of the text, allows a better understanding of its semantic structure and this leads to more accurate text mining results. Concordance helps to understand the word's exact meaning based on context as it can recognize particular instances in which words appear.

3. DATA AND METHODOLOGY

This section consists of information about the datasets used in this paper and aims at explaining the methodology employed.

3.1 Data

For the purpose of this work, datasets from HuffPost, UCI news and a dataset from TagMyNews have been used for analysis. The data in HuffPost dataset is encoded in JSON format, data from UCI dataset is in CSV format and TagMyNews dataset consists of data in the form of a file. The datasets comprise of news headlines, short description of news and few additional information such as author and title. It also contains metadata such as the source of news and date of publication.

3.2 Methodology

Methodology is the explanation of various aspects involved and it defines the relationship between several concepts, the purpose, and its working mechanisms.

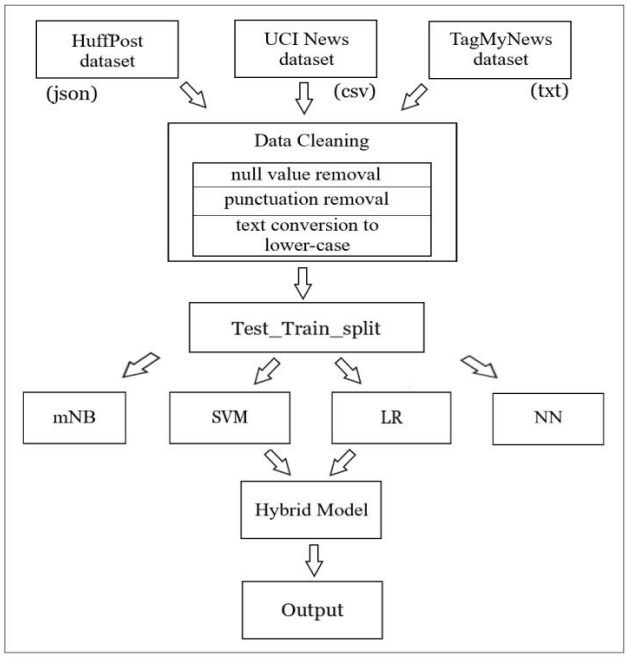


Fig-1: Design of Real Time News Headlines Classification

To understand the working of the model the process has been divided into the following parts.

A. Data Combining:

1. Data is taken from different resources and hence they're presented in different format.

2. All the datasets are converted into CSV (Comma Separated Values) from JSON and .TXT format.

3. JSON is converted into a dataframe and then to CSV.

4. TEXT format is first broken down into small text file, followed by dataframe and then to CSV format.

B. Data Cleaning:

1. Data is made even and rows containing null values are removed.
2. Punctuations are removed.
3. Whole text is converted into lower case.

C. Algorithm Analysis:

1. Dataset is divided into 2 sets, training dataset and test dataset, in the ratio 4:1.
2. Different models such as Multinomial Naïve Bayes, Support Vector Machine, Logistic Regression & Neural Network using Softmax are implemented to classify news headlines.
3. Their accuracy, precision, recall and F-1 score for different categories are calculated to analyze the working of models on different categories.

D. Creating hybrid model for real time data:

1. Based on the data generated, for each category, whichever algorithm is producing the highest accuracy for that category is selected to produce the final result for that category.
2. The model which has the highest accuracy is considered as the base model and takes care of the edge cases.
3. For instance, consider a model M1 has the highest accuracy for a certain category C4. Based on a conditional statement implemented, if output of M1 is equal to C4 then, the output will be C4 irrespective of other models classifying it in some other category.
4. Real Time data is fetched from BBC and is given as input to this hybrid model and then it is classified among the pre-defined categories.
5. Accuracy is calculated for this hybrid model.

4. EXPERIMENTAL RESULTS AND DISCUSSION

The performance of the classifiers, viz., Multinomial Naïve Bayes, Logistic Regression, Support Vector Machine and Neural Networks were evaluated by using metrics such as precision, recall, f1-score for each of the news categories. Precision: It denotes the proportion of relevant cases that are truly relevant in a dataset. Recall: It is the capability of the model to identify all the cases which are relevant in a dataset and it is also known as True Positive Rate. F1 score: It is the harmonic mean of precision and recall. All the models were trained on the same dataset and each model received different accuracy values. Term Frequency Inverse Document Frequency (TF-IDF) is used to determine the weight of terms/words in a document. It is used to assess the importance of a term in a document. TF-IDF was used with mNB and SVM classifier to see if there was an improvement in the results. Use of TF-IDF improved performance of SVM classifier, but not in the case of mNB classifier.

Table-1 gives the classification report for Multinomial Naïve Bayes (mNB) classifier with all the considered metrics. It shows the precision, recall and f1-score obtained for each of the categories.

Table-1: Classification report for mNB

	precision	recall	f1-score
business	0.87	0.85	0.86
entertainment	0.93	0.94	0.94
health	0.86	0.87	0.86
politics	0.72	0.79	0.76
science & tech	0.87	0.87	0.87
sports	0.87	0.67	0.76
world	0.68	0.64	0.66
accuracy	0.87		

Table-2 gives the classification report for Logistic Regression (LR) classifier with all the considered metrics. It shows the precision, recall and f1-score obtained for each of the categories.

Table-2: Classification report for LR

	precision	recall	f1-score
business	0.87	0.90	0.89
entertainment	0.95	0.96	0.96
health	0.88	0.91	0.89
politics	0.85	0.76	0.80
science & tech	0.90	0.91	0.91

sports	0.89	0.74	0.81
world	0.77	0.64	0.70
accuracy	0.90		

Table-3 gives the classification report for Support Vector Machine (SVM) classifier with the use of TF-IDF and this improved the performance of SVM in terms of all metrics. It shows the precision, recall and f1-score obtained for each of the categories in the table.

Table-3: Classification report for SVM

	precision	recall	f1-score
business	0.88	0.89	0.89
entertainment	0.95	0.96	0.96
health	0.88	0.90	0.89
politics	0.72	0.77	0.79
science & tech	0.90	0.90	0.90
sports	0.86	0.78	0.82
world	0.74	0.65	0.69
accuracy	0.90		

Table-4 gives the comparison between the classifiers with respect to accuracy. We can conclude from the table that Support Vector Machine (SVM) and Logistic Regression (LR) classifiers provides the highest accuracy for the considered dataset. Previous studies have shown better performance of Multinomial Naïve Bayes for text classification [7], it does not show better performance than Logistic Regression and Support Vector Machine.

Table-4: Performance of the Classifiers

Algorithm	*Accuracy*
Multinomial Naïve Bayes	86.59
Logistic Regression	89.63
Support Vector Machine	89.66
Neural Network	88.32

The models were compared with one another based on their performance in each class. The following was obtained as the results.

Fig-2 shows a graph where Multinomial Naïve Bayes (mNB) classifier and Logistic Regression (LR) classifier is compared with respect to a variety of categories. It was observed that the accuracy of Multinomial Naïve Bayes is comparatively less.

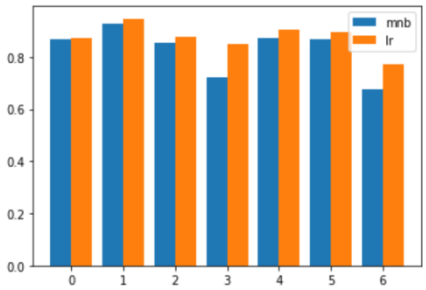


Fig-2: Comparison between mNB and LR classifiers

Fig-3 shows a graph where Support Vector Machine (SVM) classifier was compared with Neural Network (NN) classifier for a variety of categories. It was observed that the accuracy of NN was less when compared to SVM.

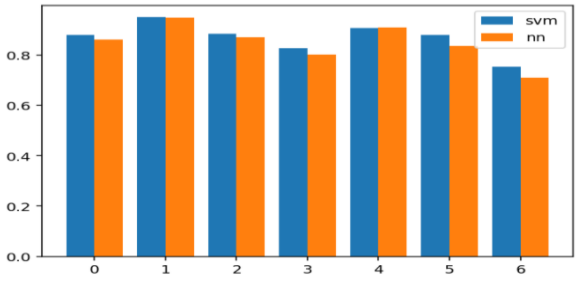
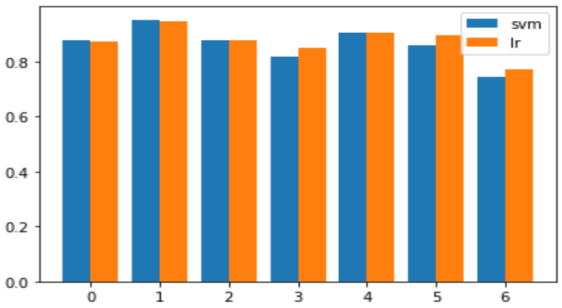


Fig-3: Comparison between SVM and NN classifiers

Fig-4 shows that Support Vector Machine (SVM) shows better accuracy in certain categories while Logistic Regression (LR) has better accuracy in others.



It was observed that, Support Vector Machine (SVM) and Logistic Regression (LR) showed the highest accuracy

when compared to the other classifiers and they were giving the highest true positive rates. As a result, a Hybrid Model was created from SVM and LR classifiers. The newly

obtained model was trained over the same training dataset, which resulted in an accuracy of 89.79 % which is an improvement of 0.13% over SVM and 0.16% over LR model. This hybrid model was used to fetch real time data and classify it based on its headline.

5. CONCLUSION

The paper focused on real time news classification on the basis of its headlines using machine learning. A system was designed to classify each news headline to its pre-defined category. Classifiers such as Multinomial Naïve Bayes, Logistic Regression Support Vector Machine, and Neural Networks were compared and evaluated. Precision, recall and F1 score were used to analyse the performance of the classifiers. The efficiency of the Support Vector Machine classifier was improved with the use of TF-IDF and this could be attributed to normalization of feature vectors and the ability of TF-IDF to find important terms/words. Performance of SVM and LR was found to be better than other models which is consistent with results of previous research [5], [11]. However, mNB classifier did not perform better, which is in contrast to the findings of previous works [7]. The reason could be because of large data set. A Hybrid Model was created from SVM and LR classifiers since they showed the highest accuracy when compared to the other models. The hybrid model was trained such that the machine is able to predict the category of the news item accurately. This approach can be used to improve the accuracy of any standalone model by combining it with different models.

REFERENCES

[1] U. Suleymanov, S. Rustamov, M. Zulfugarov, O. Orujov, N. Musayev and A. Alizade, "Empirical Study of Online News Classification Using Machine Learning Approaches," 2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT), 2018

[2] Rana, Mazhar Iqbal & Khalid, DR & Abid, Fizza & Ali, Armugh & Durrani, Mehr & Aadil, Farhan, "News Headlines Classification Using Probabilistic Approach," VAWKUM Transactions on Computer Sciences, 2015

[3] M. Ali, S. Khalid, M. I. Rana and F. Azhar, "A probabilistic framework for short text classification," 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), 2018

[4] Deb, Nabamita & Jha, Vishesh & Panjiyar, Alok & Gupta, Roshan, "A Comparative Analysis of News Categorization Using Machine Learning Approaches," International Journal of Scientific & Technology, 2020

[5] S. Yıldırım, D. Jothimani, C. Kavaklıoğlu and A. Başar, "Classification of "Hot News" for Financial Forecast Using NLP Techniques," 2018 IEEE International Conference on Big Data (Big Data), 2018

[6] Xin Liu, Gao Rujia and Song Liufu, "Internet news headlines classification method based on the N-Gram language model," 2012 International Conference on Computer Science and Information Processing (CSIP), 2012

[7] Frank E., Bouckaert R.R, "Naive Bayes for Text Classification with Unbalanced Classes," In: Fürnkranz J., Scheffer T., Spiliopoulou M. (eds) Knowledge Discovery in Databases: PKDD 2006. PKDD 2006. Lecture Notes in Computer Science, vol 4213. Springer, Berlin, Heidelberg, 2006

[8] Gurmeet Kaur, Karan Bajaj, "News Classification and Its Techniques: A Review," IOSR Journal of Computer Engineering (IOSR-JCE), 2016

[9] Jasneet Kaur, Seema Bhagla, "News Classification Using Naïve Baye's Classifier," International Journal of Advanced Research in Computer Science and Software Engineering, 2016

[10] Deshmukh, Ratnadeep & Kirange, D "Classifying News Headlines for Providing User Centered E-Newspaper Using SVM," International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 2013

[11] Joachims T, "Text categorization with Support Vector Machines: Learning with many relevant features," In: Nédellec C., Rouveirol C. (eds) Machine Learning: ECML-98. ECML 1998. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), vol 1398. Springer, Berlin, Heidelberg, 1998