International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-05, Issue-02, May 2019

Location Based News Classification using Machine Learning

[*]Shivani R Pawar, [#]Mr. M Jayashankara

[*]M.tech Student, [#]Assistant Professor, [*,#]PES College of Engineering, Mandya, India,

[*]shivanirpawar@gmail.com, [#]jayashankar.m80@gmail.com

Abstract **As we all know Web contains a vast amount of information which is gigantic and it will be changing continuously for every minute and we also know in this hectic life style it is very difficult to keep track of every news and articles that is going on. So, People are mostly concentrated on the news which is going in their nearby Environment. In this paper, we concentrate on displaying the news directing on the nearby cities and also displaying the required news articles based on few important cities. Here, we have evolved with our own web crawler to withdraw the content from the HTML pages of the articles. Random forest, Naïve Baeys and SVM classifiers are used to compute the precision and their precision is being calculated. The Machine Learning is the well known technique used for this type news classification and displaying of the news articles.**

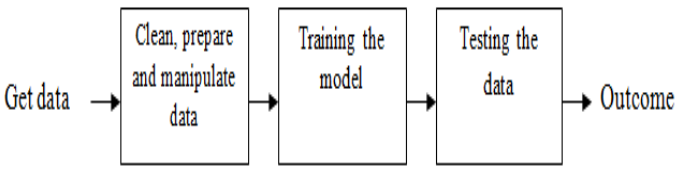
Keywords — ***Random Forest classifier, Naïve Baeys classifier, SVM classifier, Machine Learning, Natural Language Processing, Text classifiers,***

I. INTRODUCTION

The large data sets of news articles is fled out in to the world at very constant rate. It is difficult to keep track of news that is displayed everyday for a individual due to different and restless lifestyle. So people are mainly interested in online news articles and as their interest might change from individual to individual the data displayed should also be changed as the requirement and interest of a person changes. People are generally interested on the news which is going on in their instantaneous surrounding and environment. The information got might be of fewer important to a individual So, the withdrawal of expected and the applicable information for a particular person is a important task to be done. This is done in this paper mainly by concentrating a issue of interest for a particular person pertaining to that city or country or mainly to which place that person currently belongs to or is in. In this case we mainly concentrate on the geographical domain pertaining to the interest of a particular person or individual. So if we consider a example where a person wants to read a news of Mysuru but if the news displayed will be of other state or city then it will be a clumsy task to fetch a particular news. So in this case we are concentrating on displaying the news pertaining to that particular city on which a individual is interested upon.

We have used Machine learning techniques to extract the news articles based on the different classifiers and their classification. The location which we want can be of the same city, state or the country but here we are mainly focusing on extract the news of a particular city. So the news articles from the different websites of four different news papers have been taken to extract the required information. The four different newspapers are India today, The Hindu, Times of India and Indian Express.

The fundamental from of the Web page is the HTML Language. It contains various elements like comment section, bars, advertisement, news basis etc,. The text classification related here is got down with very less correctness and reliability. So to create a news articles with proper display of information and the related data to the user. The problem solving techniques should be added on to remove the efficiency and to increase the accuracy of processing the required news articles and the required classification of news based on the city. So the processing can be done by creating our own web crawler to withdraw the required news and the required webpage of the main heading of the article. For the processing to take place different operation like tokenizing the text and stemming of those tokens is done and then we remove the stop words once the stemming is done. Eventually classification is done and the trained classifiers are used to predict the output precision. Random forest, Naïve nayes and SVM classifiers are used for the classification part.



Fig(i):Steps in Machine Learning

 DOI : 10.35291/2454-9150.2019.0083

II. RELATED WORK

[1]As we all know there is huge amount of unstructured data that is being flown into the world and this unstructured will be in large scattered manner where collecting the information is very hectic task. Reading and searching such type of information is also tedious task. As we also know that the information or news these days are available through various sources such as electronic media, digital media, web media etc. In this paper they have mainly concentrated on displaying the news based on their Headlines. So in order to do that first they have gathered the information with help of RSS and then the tokenizing and retrieval of required information is done with help of KNN neighbour approach.

[2]In this paper classification of news articles is done through tf-idf approach where the large amount of unstructured data is taken and classified and pertaining to the interest of multiple-label topics for the required information to be gained. They have used the naïve baeys approach where the given problem is divided into sub problems and the different algorithmic approaches are applied for such problems. The articles here are divided as it is required like the applications or for the research purpose. The difficult part is the news articles often fall in different categories where the particular type of information to be gained becomes complicated. They have considered the 9 different fields with five years old data to be fed on and those nine different fields are Business, Sports , Entertainment, Arts, Technology, Style, Books, Home and Health. The tokenizing and stemming is done with help of tf-idf approach.

[3] This paper mainly concentrates on news articles which are fake and irrelevant and which come from irrelevant sources. The natural language processing is used in this process where the detection of fake news is done and this is detected and shown to the user. As we get the information or data from various sources like the digital media and other open source websites these data or information are checked with a help of a tool called term frequency and inverse document frequency which is also called as tf-idf approach. As we all know there are lot of information which are fake and relevant is flowing through the internet world though there are any tools which helps in detecting those like the facebook now I using few flags where it can detect the fake news but still the fake news flows even into the facebook. So to overcome such fake news and their consequences the random forest and the tf-idf type a approaches are used. The fake news everywhere in every form in the digital world detecting them is a tedious task but with the help of various approaches it can be detected and the proper news with relevant articles is displayed.

[4] As the title indicates this paper concentrates on using Support Vector Machines and Random Forests to Detect Advanced Fee Fraud Activities on Internet. The vulnerabilities of news articles are removed in this paper. The fraud activities are increasing day by day in social media platform. This can be controlled by various approaches and classifications. This paper mainly concentrates on news articles which are fake and irrelevant and which come from irrelevant sources. They have used different techniques and classifiers to extract the news articles based on the different classifiers and their classification. The location which we want can be of the same city, state or the country but here we are mainly focusing on detecting the fraud activities. So the news articles from the different websites and digital media are extracted. The data here are also got from e-mails and other digital media. So the vulnerabilities of news articles are removed in this paper. The SVM and random forest classifiers plays a very important role in this paper. So in order to do that first they have gathered the information with help of RSS and then the tokenizing and retrieval of required information is done with help of KNN neighbour approach and the fraud detecting activities are done with other classifier techniques.

[5] Huge amount of information is produced everyday with the rapid growth of world wide web and electronic information devices. So in this paper they have used the RSS feeds to display and extract the information from various sources. The information are scattered in different forms across various media. They are aiming at getting the information from four different places and their criteria's. This would save a lot of time where the user can mainly concentrate on a particular data of information. This paper mainly concentrates on the feeds taken from various websites and these information and displayed as required by the user.

III. METHODOLOGY

Based on the required content the output class is applied to the news articles and this is our main approach. The entire flow of process is shown in the fig1. The process starts with the data retrieval process where the data is retrieved and the having our self developed web crawler which is used to extract the data. Then the process flown is divided into the train data and the test data where the train data contains the 80% and the test data contains 20% of the data retrieved. Text data processing methods are applied to the train data which is in turn given as input to the classifier for training. Even for the test data the text data processing methods are applied where the it is given as input to the trained classifier and then the precision of output classes are got by applying the algorithms. The evaluation for accuracy is done by applying some performance metrices. This is explained in detail in further sections.

 DOI : 10.35291/2454-9150.2019.0083

(i) Data Retrieval

This is the first phase of the process where the news articles are collected from various sources of websites. It contains three different steps where the first step is Parsing the RSS feeds to get the URL's during this phase the URL's from different websites is been retrieved where the data is collected from these websites. The second step is to collect the URL's in file where it can used further for processing so this is saved and filed for future use. The last is to Extract the articles using URL's here fetching of information is done from various websites through these URL's. each news articles is extracted by visiting every URL's and their websites. Web crawler helps in extracting and displaying the news through HTML page for different news papers and those are India today, The Hindu, Times of India and Indian Express.

(ii) Text Data Processing

Text data processing contains four different steps. It involves the pre-processing of the extracted data. In this phase first the tokenizing of the fetched data is done where the string of characters are divided as required only the required form of string is saved and the unwanted string of characters are removed. Stemming is the next step where the words in the articles are reduced to their declension form. This is followed by stop word removal, the stop words are those which are of no importance in fetching and reading the required data so such articles are removed from the fetched data.

(iii) Training the Classifier

As the name indicated this phase is used to train the classifier that is a set of fetches data as trained and test data. At first the pre-processing of modules is done at further steps including three different steps. First we have the text pre-processing module where the relevant part of the articles are collected. The input to the classifier is done in two different forms of vectors: counter vector and hash vector here the vectorization of the articles and correspondence labels are applied to the each article which is fetched. Once the vectorization is done it is fed to the classifiers. Finally the vectorizer and the classifier is stored in a file for further processing.

(iv) Testing Classifier

The file containing the classifier which is stored and the counter object is loaded for test classifier. Here the testing of fetched data is done so the stored classifier is trained and the testing data is fed into it. The accuracy is determined in this module where the precision is calculated for the respective algorithms applied. The trained classifier classifies the performance of the precision. Once the testing is done the accuracy of the classifier is noted on different performance metrices like Precision, Recall and F1 score. The classification is done on the basis of city.

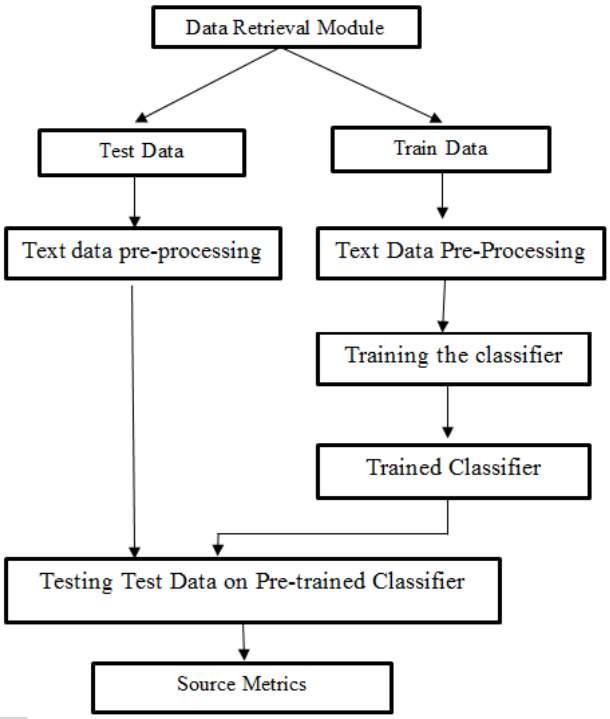


Fig1: Flow chart of the Process

IV. RESULTS

The final outcome gives the precision, recall and F1 score. The score is considered as the primary performance measure where the better performance of different algorithms is noted. It is mainly divided into precision and Recall

(a) Precision

It is the ratio of the number of articles which are true positive that is which are correct to the total number articles classified and predicted having a particular category. It is mainly classified as positive and negative value where her it is called as positive predictive value.

It can be defined with following formula:

$P = \frac{m}{m+n}$ here, m stands for True positive and n stands for false positive

(b) Recall

It is the ratio of number of articles which are true positive to the number of actual articles with a particular category. Even here it consists of positive and negative values.

It can be defined as follows:

$R = \frac{m}{m+t}$ here, m stands for True positive and t stands for false positive.

the end result gives us the process time and the predicted results with the help of bar graph as shown in the fig2 and

 DOI : 10.35291/2454-9150.2019.0083

fig3 where the processing time of different algorithms and their predicted results is shown respectively.

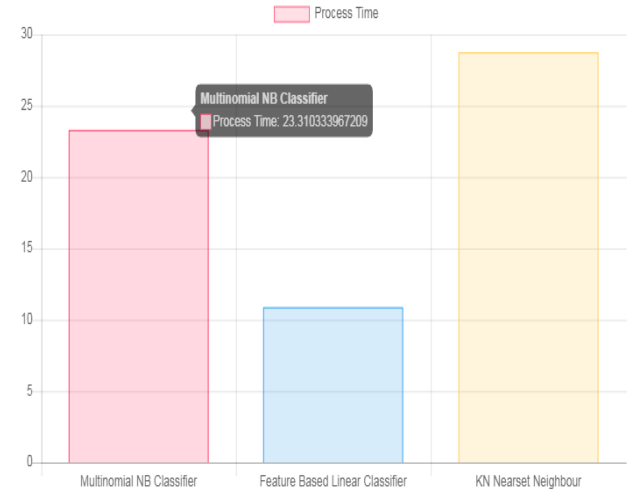


Fig2: Process time

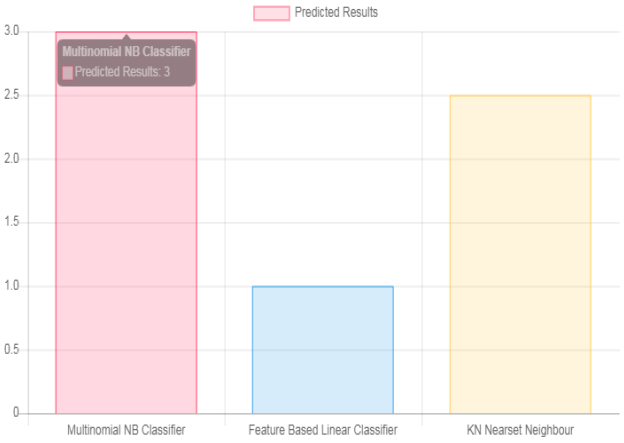


Fig3: Predicted results

V. CONCLUSION

The goal of the project is to make use of different techniques of machine learning to classify news articles based on their different locations. We have used different algorithms like Naïve Bayes, Support vector machine and Random forest. The accuracy of the outcome of these classifiers is measured using performance matrix and finally best classifier will be selected. The proposed system can be used as a part of more complex news article classification systems. The neural networks can be used to have the more precise outcomes and the delay can be decreased.

REFERENCES

[1] MI Rana, S Khalid, MU Akbar- News classification based on their headlines: A review Multi-Topic Conference (INMIC) …, 2014 - ieeexplore.ieee.org.

[2] Z CHASE, N Genain, O Karniol- Learning Multi-Label Topic Classification of News Articles…Tambour - 2014 cs229.stanford.edu..

[3] Shlok Gilda- Evaluating Machine Learning Algorithms for Fake News Detection….,2017 IEEE 15th Student Conference on Research and Development

[4] Abiodun. Modupe, Oludayo. O. Olugbara and Sunday. O. Ojo- Exploring Support Vector Machines and Random Forests to Detect Advanced Fee Fraud Activities on Internet…., 2011 11th IEEE International Conference on Data Mining Workshops.

[5] B. Pendharkar, P. Ambekar,P. Godbole, S. Joshi, and S. Abhyankar, Topic categorization of rss news feeds,"Group vol. 4, p. 1, 2007.

[6] J. Davis and M. Goadrich, The relationship between precision recall and ROC curves, in Proceedings of the 23rd International Conference on Machine Learning, ser. ICML 06. NewYork, NY, USA: ACM, 2006, pp. 233240 [Online].

[7] T. Landgrebe, P. Paclk, R. Duin, and A. Bradley, Precision-recall operating characteristic (P-ROC) curves in imprecise environments, in Proceedings of ICPR, 2006.

[8] H. a. K. S. Yu, "SVM tutorial: Classification, regression, and ranking", Handbook of Natural Computing 2009.

[9] B. Pendharkar, P. Ambekar,P. Godbole, S. Joshi, and S. Abhyankar,"Topic categorization of rss news feeds," Group vol. 4, p. 1, 2007.

[10] Mingyong Liu, Jiangang Yang, "An improvement of TFIDF weighting in text categorization," International Conference on Computer Technology and Science, IPCSIT vol. 47, IACSIT Press, Singapore, 2012.

[11] M. Ikonomakis, S. Kotsiantis, V. Tampakas. Text Classification Using Machine Learning Techniques. WSEAS TRANSACTIONS on COMPUTERS, Issue 8, Volume 4, 966-974, 2005.

[12] Aggarwal, C.C., Zhai, C.: A survey of text clustering algorithms. In:Mining Text Data, pp. 77–128. Springer (2012).

[13] Kim, S.B., Han, K.S., Rim, H.C., Myaeng, S.H.: Some Effective Techniques for Na¨ıve Bayes Text Classification. IEEE Trans. Knowl. Data Eng. 18(11), 1457–1466 (2006).

[14] S. Ting, W. Ip &A. H. Tsang, "Is Naive Bayes a good classifier for document classification?," International Journal of Software Engineering and Its Applications, vol 5 no 3, 2011

[15] R. D. Goyal, "Knowledge based neural network for text classification, "in proc. of the IEEE international conference on Granular Computing,pp. 542-547, Nov. 2007.

[16] A. M. Mahmood, N. Satuluri, and M. R. Kuppa, "An Overview of Recent and Traditional Decision Tree Classifiers in Machine Learning.", International Journal of Research and Reviews in Ad Hoc Networks, Vol. 1, No.1, 2011.

[17] Podgorelec V, Kokol P, Stiglic B, Rozman I: Decision trees: An overview and their use in medicine.Journal of Medical Systems. 2002,26:445–463.

[18] Vandana Korde,C Namrata Mahender "Text classification and classifier:A survey" International Journal of Artificial Intelligence & Applications. 2012.

 DOI : 10.35291/2454-9150.2019.0083