International Journal of Scientific Research & Engineering Trends
Volume 6, Issue 2, Mar-Apr-2020, ISSN (Online): 2395-566X

Machine Learning Approach to Classify News Articles Based on Location

Shrina Sachin, Monika Kar, Swapnali Shewale, Prof. Shweta Barshe
Department of Computer Engineering
Bharati Vidyapeeth's College of Engineering, Navi-Mumbai, Maharashtra, India

***Abstract* – In this hectic lifestyle, we need to classify news as per user requirements. News are important and things are constantly changing around the world. In this paper, we will be classifying the news articles based on cities and providing set of cities specific news. We have developed our own web crawler for content extraction from the HTML pages of news articles. Random Forests, Naive Bayes and SVM classifiers algorithms are used. Machine learning techniques are used to achieve our goal and helps to classy news articles.**

***Keywords*– Support Vector Machine, Naive Bayes classifier, SVM Classifier, Multinomial Naive Bayes.**

I. INTRODUCTION

In this research, we have implemented machine learning techniques to classify news articles based on location. The location can be a city, state, country, etc. but we have examined the results based on cities. The news articles from various newspaper websites like Hindustan Times, Times of India etc. are extracted to form our dataset.
In this text classification we will design web crawler in order to check the website and extract the news articles from the webpage. Next step would be tokenization of words to its root form for example environmentalist gets tokenized to environment. Tokenization would be followed by stop words removal. In the next step classification is performed and classifiers are trained in order to obtain output class. Classification have been performed with help of algorithms such as Random Forest Classifier, Support Vector Machine and Multinomial Naive Bayes.

1. Random Forest Classifier

Random Forest is a supervised classification algorithm. It uses a number of uncorrelated decision tree classifiers and fits them on various sub samples of dataset. It is used in outlier detection and replacing missing data. It is scalable as it can run on large data sets.

2. Naive Bayes classifier

Naive Bayes classifier is a popular method for text classification. It is based on probabilistic technique of classification which derives its roots from Bayes Theorem. It is based on the assumption of independence between the various features. It is a scalable classifier and can run efficiently with large data sets.

3. SVM Classifier

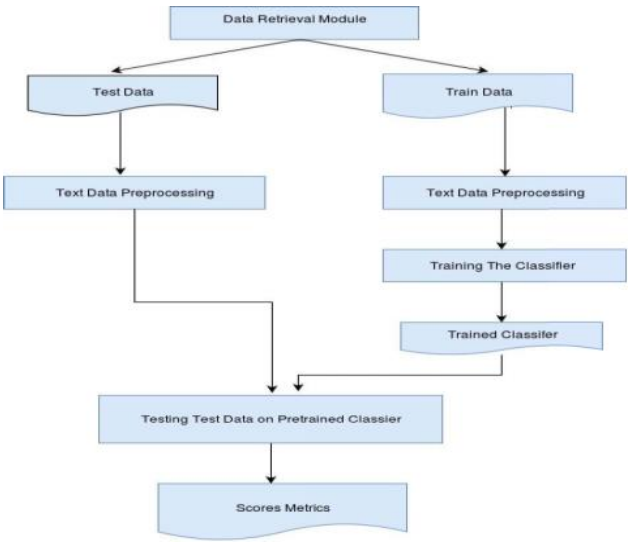
SVM (Support Vector Machine) is a supervised classification algorithm. It has a training set and labels associated with it. After training, if a test data is fed in, the model assigns it to one category or the other. It performs well with linear classification. It constructs a hyper- plane for classification. The hyper-plane is chosen such that the distance between the nearest data point on either side is maximized

II. LITERATURE REVIEW

Existing System shows all news at a time, in today's society the world population get their news on their electronic devices. Many of the major newspapers have articles displayed in a flow. Many newspapers have divided their articles into categories. The problem is then that a user would have to go to the different pages to read about different categories would be simplified.

III. METHODOLOGY

- **Flow Chart of Process**



- **Working Principle**

The aim of the project is to fetch news according to the user-oriented locations. It will be achieved with the help of machine learning algorithms such as Random Forests, Naive Bayes and SVM classifiers. In this project natural level processing is also performed upon the retrieved texts.

The texts are divided into test data and train data and later on data preprocessing is done. Train data consist of 80% and test data consist of 20% of the dataset. Train data acts as input for the classifier and test data acts as an input to the trained classifier which predicts the output class of the news articles.

IV. CONCLUSION

Thus, we have investigated the possibility to use machine learning algorithms to classify the news articles based on cities. The experiments show that this problem can be successfully solved by using various Classifiers such as Naive Bayes, Support vector Machines and Random Forest. Random Forest has outperformed the other classifiers. Naive Bayes has performed well too and Support Vector machine is at the bottom in terms of the performance metrics used in our approach.

In this project we have shown that we can indeed use a text classifier to select interesting articles from a small dataset based on personal opinion such as user location-oriented news.

V. FUTURE ENHANCEMENT

If we could continue this experiment, we would want to conduct an experiment on what the best way is to find out if a user thought an article was interesting or not. We would like to compare how the classification confidence change depending on how the classification of articles are gathered from the user. For example; the time spent on an article could be used as an indication on whether the user thought the article was interesting or not. Another could be that the user could classify the articles themselves with the use of like/dislike buttons. To make the results more relevant it would be interesting to research if a classifier trained on one newspaper's articles could be used to classify articles from other newspapers. We would also like to do further testing on datasets with larger sizes to see if we can further improve
the results from this project.

REFERENCES

[1]. Data Mining Concepts, Authors - Sung- Shiou Shen, Tsai-Hua Kang, bhen-Ho Lin & Wei Chien, Proceedings of the 2017 IEEE International Conference on Applied System Innovation.

[2]. SVM Tutorials: Classification, Ranking and Regression. Authors – Chadramohan Sudar, Arjun SK, Deepth L R. Dated – 2017 IEEE explorer.

[3]. Information Retreival and Techniques. Authors– Rajkumar Janakiraman, Sandeep Kumar, Sheng Zhang, Terence Sim. Proceedings of the Seventh IEEE workshop on application of computer vision.

[4]. Automated Online News Classifications. Authors – Towseef Akram, Vakeel Ahmad, Israrul Haq, Monisa Nazir. International Journal of Computer Science and Mobile Computing. Vol-6, Issue- 6, June 2017.