Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 3, No. 1, March : 2024

DETAILED INVESTIGATION OF TEXT CLASSIFICATION AND CLUSTERING OF TWITTER DATA FOR BUSINESS ANALYTICS

Ziya Fatima M Tech Scholar, Department of Computer Science Engineering, SVIET, Banur, Punjab, India
Er. Prince Sood Assistant Professor, Department of Computer Science Engineering, SVIET, Banur, Punjab, India

INTRODUCTION:

Social media has significantly transformed us and the way that we associate with the world. Latest research shows that a large number of individuals are utilizing social media like Facebook and Twitter for the purpose of finding and sharing data, making companions and engaging themselves. An advantage of social systems is that they can show the opinions of people which can be used for the purpose of investigating the present situation and anticipating the future. There are different means of user interaction can be via exchange of messages, exchange of files and pictures, exchange of videos and audio. As of late, investigation into social media has been done using the information gathered from online stages like Facebook, Twitter, LinkedIn and Flickr in various fields like business, governmental issues, education and medicine. The ability to gather such information is significant in social media research and led to the formulation of a new field called computational social science.

LITERATURE REVIEW:

In paper [1], Kader has analyzed sentiment analysis of twitter using naïve bayes classification algorithm. Collected data from twitter and then preprocessing techniques are applied to clean the data. Then analysis is done based on some improvisation in naïve bayes for NLP processing and improved the accuracy rate by 5%. The map reduce framework has been used in order to handle big data. The reason being that it is reliable, simple and scalable. In research paper [3] has explained that the sentiment analysis is most active area of research in big data. Big data characteristics in sentiment context is shown in the paper. Analysis of the data is done in Hadoop based on the big data characteristics and SA application in various aspects are analyzed. In paper [2], Goyal has discussed in the paper regarding big data analytics for demonetization data collected from Hadoop platforms pig and hive. The analysis has been reviewed and showed in the paper. Person’s review analysis has been surveyed from data collected and some analysis based on the collected demonization data is done. The data that has been analyzed has been generated from the views of the people on various social media platform. In paper [9], it is stated that there are a lot of opinions expressed by people on a product in the form of a review. These reviews can be used to gauge the sentiment of the people or the customers towards a product. On analyzing the same a company can understand the needs of a customer better and also see which product is selling in which are more. The procedure for the analysis of sentiments is followed like collection of data and tokenizing it and POS tagging is used in accordance with a dictionary of Sentiword for gauging the sentiments. In paper [10] various classification algorithms can be used for sentiment analysis. Decision trees have fast fitting speed and fast prediction speed, but have low accuracy. Naïve bayes has high accuracy but has slow prediction speeds and consumes much time in training. This paper has described the various algorithms that are used for big data analytics. The comparisons have been made between the various algorithms and also the applications of the same have been described. In paper [6], the author has used the Naïve Bayes Algorithm for the purpose of classification of dataset in order to derive information that is of great help in checking the sentiments of the people towards a certain issue. The MapReduce of Hadoop technology has been used. Preprocessing of the data set and tokenization are done after the data has been extracted from twitter. In the two phases of map reduce, the first phase i.e. mapping is used for token generation of the overall post or tweet. In the second phase i.e. reducing classes for categorization are created and emoticons are converted to text. The efficiency is increased. In paper [7], Twitter’s API is used to derive data from twitter. Then the data undergoes some processing. First removal of stop words is done. Then the tokens are changed into a structured form as they are mostly in unstructured form. Then emoticons are also translated for higher accuracy. Then Map-Reduce is used to find sentiment of each word and the sum gives the overall sentiment of the tweet. In paper [8] first map-reduce, a sentence is detected and stop words, hashtags etc. are removed. Then we search for words which represent features and are then clustered. Then Open NLP is used for POS Tagging. Phrase removal is done before stop word removal. In second map-reduce, a sentiwordnet dictionary is used, scores are given to words, which is averaged then to get overall value. In paper [5], has examined the sentiment of youngsters regarding the floods in Chennai in 2016. They used flume to get data from twitter and applying Naïve Bayes algorithm. They developed a dictionary to compare the tweets with and get a sentiment score. In paper [4] various methods and concepts of sentiment analysis on big data are showed. It has two approaches- Lexicon based and Machine Learning based. The machine learning approach is more popular and uses various supervised and unsupervised learning algorithms.

RESEARCH METHODOLOGY USED:

Data Collection: Data is available in abundance via the internet today. In today’s competitive world, none of the data can be considered as not useful. It is a resource that if utilized in a proper manner can provide great results and benefits to a company. The data in this work has been collected from twitter using the twitter API. Twitter data has been used as it is mostly text.

Filtering/Prior Processing: The data that is collected from twitter is not in a form which can directly produce results. This data this is required to be filtered for removal of parts which are of no use to generate results. This includes removal of hashtags, stop words, marks of punctuation etc.

Feature Optimization: This process helps in the construction of a model. It helps in reducing the training time of the designed model. This works on the concept that there might be some parts in a dataset that are repeated over and over again or that might not show any relevance to the topic at hand. Minimization of error rate is also a part of this process.

Classification Using SVM: Classification is important when it comes to categorization of text or recognizing the tone, classification of images and so on. In this work support vector machine (SVM) has been used. It is a powerful tool when analyzing data. Several metrics are used for measuring the performance. The used metrics are accuracy, precision, recall.

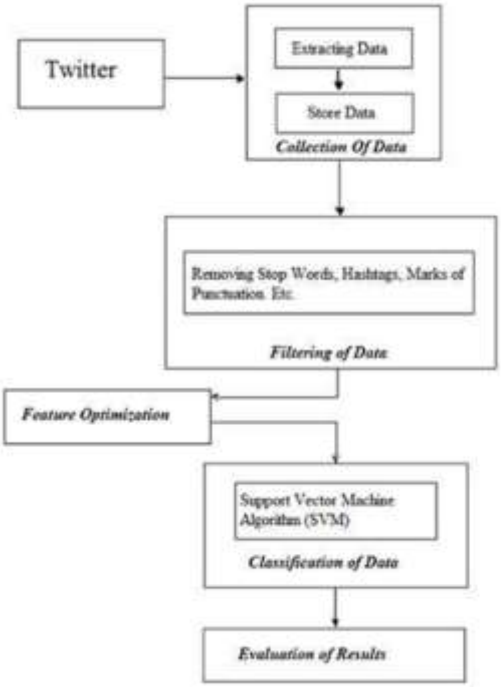


Figure 1: Flowchart for the Methodology Proposed.

SIMULATED RESULTS AND DISCUSSION:

The first step is to load the dataset in order to perform the analysis on the same. The dataset is loaded in order to remove the stop words so that result accuracy is improved. The sample data taken has been generated from twitter. The tweets of many users have been taken which contain any information related to tax. This extracted data is then used to perform the analysis so as to understand the opinions of various people on the related issue or topic. The sentiment are further grouped on the polarity using feature optimization.

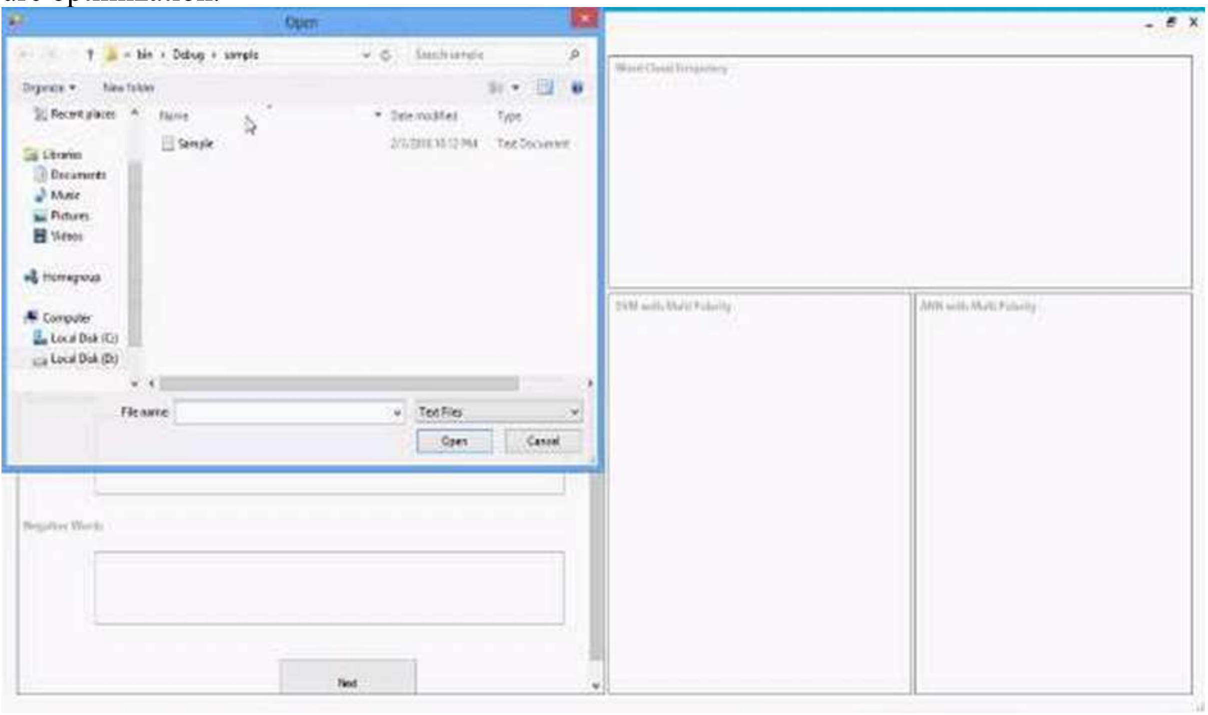
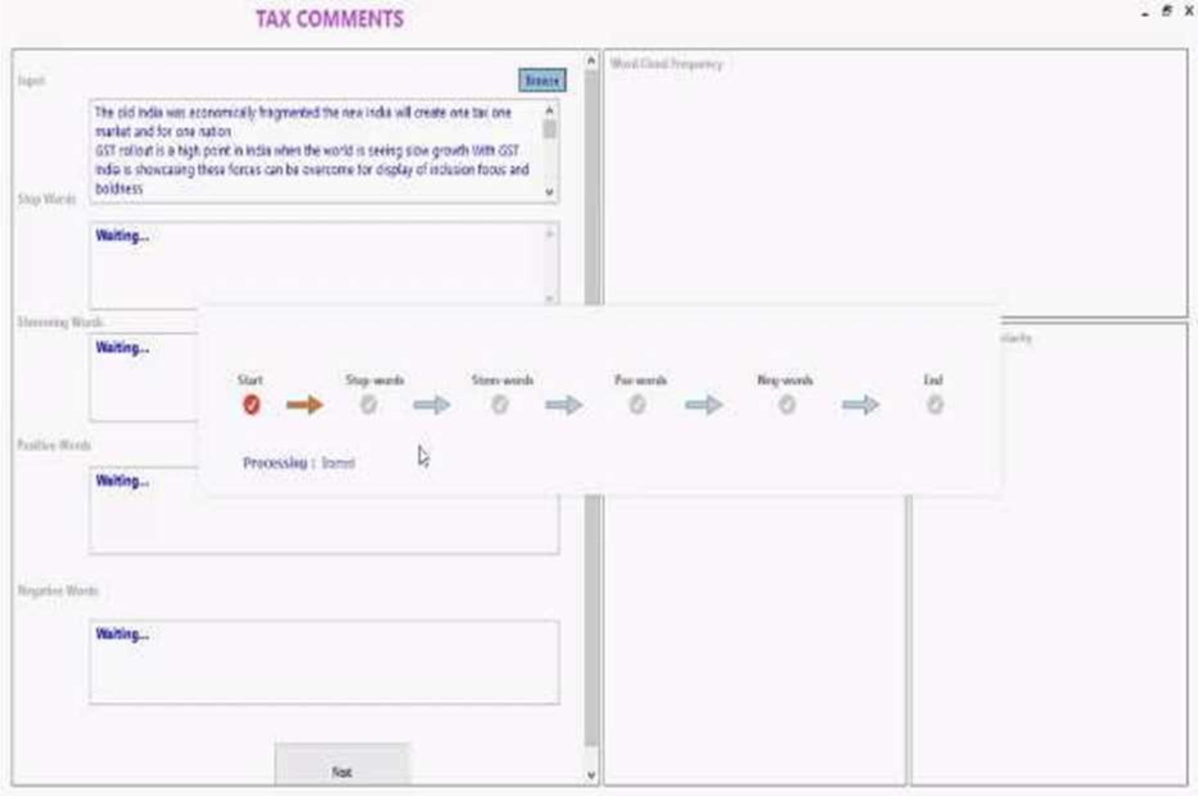


Figure 2: Loading the dataset



The old india was economically fragmented the new india will create one tax one market and for one nation
GST rollout is a high point in india when the world is seeing slow growth With GST india is showcasing these forces can be overcome for display of inclusion focus and boldness

Waiting...

Waiting...

Processing : Started

Waiting...

Waiting...

Figure 3: Identifying the words.

After loading the dataset the process of opinion mining or what we call sentiment analysis starts. The stop words and stem words are identified followed by positive and negative words. The time taken depends on the size of the dataset. Feature selection helps in achieving a better performance and leads to more accuracy. Processing the dataset has been done using SVM and feature selection and ANN. The polarity results that follow clearly show the results based on polarity of the sentiments expressed by the people. And SVM performs better.

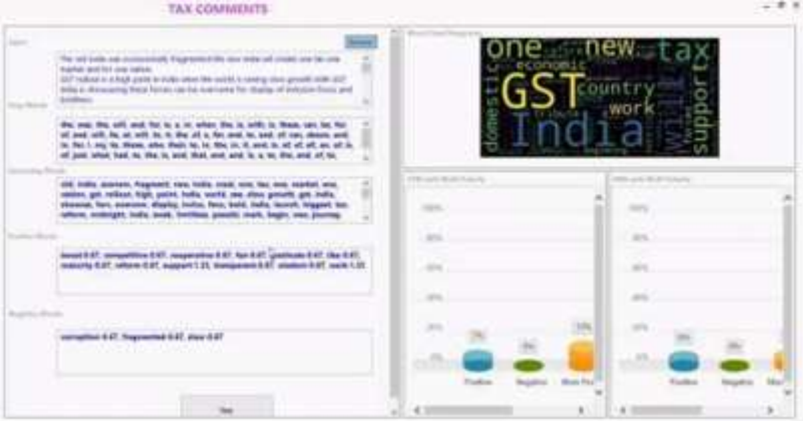


Figure 4: Results based on SVM and ANN.

Figure 5: Evaluation of Results.

The results are further evaluated. A comparison has been made between the accuracy of the predicted results and the actual results generated. Also, a Comparison between the performances of SVM and ANN has been made. The results are displayed in the figure that follows.

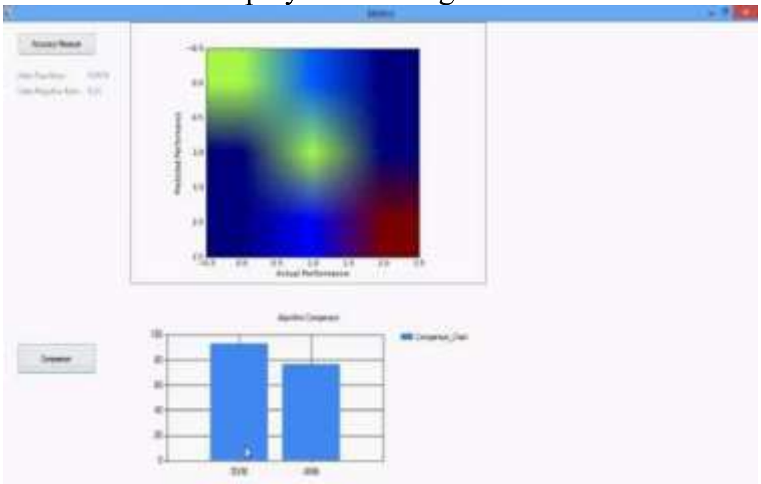


Figure 6: Accuracy Measure and Comparison.

A text mining approach thus using SVM algorithm has been done with feature selection. This is done with a view of getting the SVM environment work in a much way that is to perform better. The proposed technique performs much better than the existing technique.

CONCLUSION AND FUTURE SCOPE:

The outcome of evidence-based decision making contributes to the improvement of a brand. Having a text analysis of tax data allows effective quality management. With sentiment analysis, companies can now strategically reposition their businesses according to customers' sentiments.

This work provided an introduction and rationale behind the value of text analytics of Twitter data to businesses. The work presented in this dissertation also discussed several related work in sentiment analysis for business applications. Importantly, it demonstrated a practical application of text classification and clustering of Twitter data, and revealed ways on how to analyze these to gain business insights. Although the classification accuracy rate for this experiment is already acceptable in this application domain.

It is suggested that future work needs to increase the accuracy of the classification model by improving data preparation and experimenting with other classification algorithms. Also, tweets on some different issues can be taken into consideration for more analysis of sentiments. Also, some deep learning techniques to use to improve the prediction accuracy and more optimize the results. The same approach can be used with emotion mining also in which number of class parameters are more than three.

REFRENCES:

[1] Mariam Khader, Arafat Awajan, Ghazi Al-Naymat, "The Effects of Natural Language Processing on Big Data Analysis: Sentiment Analysis Case Study", International Arab Conference on Information Technology (ACIT), 2018.

[2] MalvikaGoyal and Anuranjana, "Demonetization- Twitter Data Analysis using Big Data & Hadoop", Amity International Conference on Artificial Intelligence (AICAI), 2019.

[3] Imane El Alaoui, Youssef Gahi, RochdiMessoussi, "Full Consideration of Big Data Characteristics in Sentiment Analysis Context", IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), 2019.

[4] M. Edison, A. Aloysius, "Concepts and Methods of Sentiment Analysis on Big Data", International Journal of Innovative Research in Science Engineering and Technology, Vol. 5, Issue 9, September 2016, pp. 16288-16296.

[5] Priya. V, S DivyaVandana, "Chennai Rains Sentiment-An Analysis Of Opinion About Youngsters Reflected In Tweets Using Hadoop", International Journal of Pharmacy & Technology, Sep- 2016, Vol. 8, Issue No.3, pp. 16172-16180.

[6] Huma Pandey and Shikha Pandey, "Sentiment Analysis on Twitter Data-set using Naive Bayes Algorithm", IEEE, 2nd International Conference on Applied and Theoretical Computing and Communication Technology, 2016, pp. 416-419.

[7] DivyaSehgal and Dr. Ambuj Kumar Agarwal, "Sentiment Analysis of Big Data Applications using Twitter Data with the Help of HADOOP Framework", IEEE, 5th International Conference on System Modelling & Advancement in Research Trends, 25th -27th November, 2016, pp. 251-255.

[8] Jalpa Mehta, JayeshPatil, RuteshPatil, MansiSomani and SheelVarma, "Sentiment Analysis on Product Reviews using Hadoop", International Journal of Computer Applications Volume 142 – No.11, May 2016, pp. 38-41

[9] Dr. U Ravi Babu, "Sentiment Analysis of reviews for E-Shopping Websites", International Journal of Engineering and Computer Science, Volume 6 Issue 1 Jan. 2017, pp. 19965-19968.

[10] HardiRajnikantThakor, "A Survey Paper on Classification Algorithms in Big Data", International Journal of Research Culture Society,
Volume 1, Issue 3, May 2017, pp. 21 -27

[11] Shawn Graham, Ian Milligan and Scott B. Weingart "Exploring Big Historical Data - The Historian's Macroscope", London: Imperial College Press, 2016.