JCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue-15, May 2019 E-ISSN: 2347-2693

Sentiment Analysis of Twitter Data Using Text Classification And Clustering

Kumari Ashmita Sinha[1*], Maanasa VKP [2], Nikhitha SP [3], Ramya S [4], Mangala CN[5]

[1,2,3,4,5]Department of Computer Science, East West Institute of Technology, Bengaluru, India

DOI: https://doi.org/10.26438/ijcse/v7si15.6972 | Available online at: www.ijcseonline.org

Abstract—Information and data has always been consideredas the lifeline to run any business. The stability of an organization, the growth of the business, the profit gained or the loss suffered, all these factors depend on the information gained about the market trends and also most importantly public opinion. The sentiments of the people are therefore considered as the most crucial data that the organizations use in order to take effective decisions with minimum risk.In this paper we gather data from a type of social media that is twitter. This is because the public nowadays express their opinions and their feedbacks largely to through social media. In this paper we attempt to perform sentiment analysis using text classification by Naïve Bayes and text Clustering by K-means.

Keywords-Twitter, Sentiment Analysis,Social Media,Naïve Bayes, K-means

I. INTRODUCTION

In the society that we live in, different people have different perspective on how they view things. Society is a large pool of diverse opinions and sentiments. For any business, it is vital to know the trends in the market. It is important to know the public opinion or public sentiments about the product by obtaining their feedback. The traditional or the conventional way of gathering feedback by manual survey is not much efficient in present day. By the old ways it is not possible to get more quantity of data. So, the data mining is done by gathering data from the source where people freely express their opinions that is social media.

From past few decades social media and the number of users have been rising exponentially. Social media like twitter has large number of users so in turn produces enormous sets of data. Therefore, the social media twitter has become one of the potential sources for extracting raw data that reflects public opinion. Such a process of extracting data and analysing it using different techniques to get to know positivity or negativity is called sentiment analysis.

Sentiment Analysis also called opinion mining or emotion AI uses natural language processing ,also text classification which is to apply on texts to find analyse the polarity. The outcome of this process is graduated on scales of accuracy and precision. Almost every time we obtain a accuracy around 80% which means the system works as well as humans that is it thinks like humans and takes decision in the same way. If a system produces 70% accuracy then it's destined to give best results. Almost all companies depend on customer's satisfaction for whom sentiment analysis acts a backbone.

In this paper, we have implemented sentiment analysis in python since it is a source of huge number of libraries and we have used two algorithms *Naive Bayes and K-means clustering* frommachine learning. Machine leaning is a major part of Artificial Intelligence focusing on making machines learn from their previous experience, experiments and work without explicitly programming.

II. RELATED WORK

"Sentiment Analysis and Opinion Mining on Online Customer Review "
The opinion mining is very much essential in e-commerce websites, furthermore advantageous with individual. An ever increasing amount of results are stored in the web as well as the amount of people would acquiring items from web are increasing. As a result, the users' reviews or posts are increasing day by day. The reviews toward shipper sites express their feeling. Any organization for example, web forums, discourse groups, blogs etc., there will be an extensive add up for information. Records identified with items on the Web, which are functional to both makers and clients. The process of finding user opinion about the topic or product or problem is called as opinion mining. It can also be defined as the process of automatic extraction of knowledge by means of opinions expressed by the user who is currently using the product about some product is called as opinion mining. Analysing the emotions from the extracted opinions is defined as Sentiment Analysis. The goal of opinion mining and Sentiment Analysis is to make computer able to recognize and express emotion. This work concentrates on mining reviews from the websites like Amazon, which allows user to freely write the view. It automatically extracts

the reviews from the website. It also uses algorithm such as Naïve Bayes classifier, Logistic Regression and SentiWordNet algorithm to classify the review as positive and negative review. At the end we have used quality metric parameters to measure the performance of each algorithm.

The advantage in this paper was that,the best classifier among three algorithms for text classification can be determined. And the disadvantage in this paper was that we cannot focus on mining reviews from multiple website the technique used here don’t allow us to mine the reviews from different websites.

"Sem Eval-2017 task 4: Sentiment analysis in Twitter"

This paper describes the fifth year of the Sentiment Analysis in Twitter task. SemEval-2017 Task 4 continues with a rerun of the subtasks of SemEval-2016 Task 4, which include identifying the overall sentiment of the tweet, sentiment towards a topic with classification on a twopoint and on a five-point ordinal scale, and quantification of the distribution of sentiment towards a topic across a number of tweets: again on a two-point and on a five-point ordinal scale. Compared to 2016, we made two changes: (i) we introduced a new language, Arabic, for all subtasks, and (ii) we made available information from the profiles of the Twitter users who posted the target tweets. The task continues to be very popular, with a total of 48 teams participating this year.

The advantage in this paper was it can Identifies overall sentiment of tweets. And the disadvantage of this paper was we cannot approach the irony and emotional detection sentiments.

"Sem Eval-2016 task 5: Aspect based sentiment analysis"

This paper describes the SemEval 2016 shared task on Aspect Based Sentiment Analysis (ABSA), a continuation of the respective tasks of 2014 and 2015. In its third year, the task provided 19 training and 20 testing datasets for 8 languages and 7 domains, as well as a common evaluation procedure. From these datasets, 25 were for sentence-level and 14 for text-level ABSA; the latter was introduced for the first time as a subtask in SemEval. The task attracted 245 submissions from 29 teams.

In its third year, the Sem-Eval ABSA task provided 19 training and 20 testing datasets, from 7 domains and 8 languages, attracting 245 submissions from 29 teams. The use of the same annotation guidelines for domains addressed in different languages gives the opportunity to experiment also with crosslingual or language-agnostic approaches. In addition, SE-ABSA16 included for the first time a text level subtask. Future work will address the creation of datasets in more languages and domains and the enrichment of the annotation schemas with other types of SA-related information like topics, events and figures of speech (e.g., irony, metaphor).

"Enhanced Twitter Sentiment Analysis by Using Feature Selection and Combination"

Tweet sentiment analysis is an important research topic. An accurate and timely analysis report could give good indications on the general public's opinions. After reviewing the current research, we identify the need of effective and efficient methods to conduct tweet sentiment analysis. This paper aims to achieve a high level of performance for classifying tweets with sentiment information. We propose a feasible solution which improves the level of accuracy with good time efficiency. Specifically, we develop a novel feature combination scheme which utilizes the sentiment lexicons and the extracted tweet unigrams of high information gain. We evaluate the performance of six popular machine learning classifiers among which the Naive Bayes Multinomial (NBM) classifier achieves the accuracy rate of 84.60% and takes a few minutes to complete classifying thousands of tweets.

The advantage in this paper was that we can evaluate performance of different classifiers to find best choice for tweet. And the disadvantage of this paper is that we have to manually label the tweets this paper cannot gives permission to label the tweets automatically.

"Lexicon-enhanced LSTM with Attention for General Sentiment Analysis"

LSTMs have gained good performance in sentiment analysis tasks. The general method is to use LSTMs to combine word embeddings for text representation. However, word embeddings carry more semantic information rather than sentiment information. Only using word embeddings to represent words is inaccurate in sentiment analysis tasks. To solve the problem, we propose a lexicon-enhanced LSTM model. The model first uses sentiment lexicon as an extra information pre-training a word sentiment classifier and then get the sentiment embeddings of words including the words not in the lexicon. Combining the sentiment embedding and its word embedding can make word representation more accurate. Furthermore, we define a new method to find the attention vector in general sentiment analysis without a target, which can improve the LSTM ability in capturing global sentiment information. The results of experiments on English and Chinese datasets show that our models have comparative or better results than the existing models.

The main advantage in this paper was that we use LTSM which has strong ability in modelling short sequences. And the disadvantage in this paper is that we need more lexicon resources for better accuracy which ranges for high cost thus making it expensive.

III. METHODOLOGY

Sequence of the project:

This project consists of two major parts:First, it is the data extraction part from social media.Here, we have considered data from twitter.

Second, it is the data analysing part where we use two machine learning algorithms to classify and cluster the data and ultimately perform sentiment analysis.

The data we extract here are the tweets from twitter which can be extracted via the twitter development website using certain API(Application Programming Interface) like tweepy, pandas, etc. This data will be fed to our classification model and our clustering model.

In this project, firstly the user authorization and user authentication will take where the user creates an account by registering and then can later login. Then once after logging in, he or she can enter the keyword(Ex: a product name) and enter the "ok" button. The tweets containing that keyword will be extracted and will be stored in an excel sheet. The tweets will be later subjected to pre-processing. This is where cleaning of the data will take place to eliminate unnecessary information(like removing special characters).Then the frequency of each and every word will be calculated and will be stored in the from of word-bag. We should fed this to our model to do the below process.

Then after classification and clustering we get the probability of negativity or positivity.

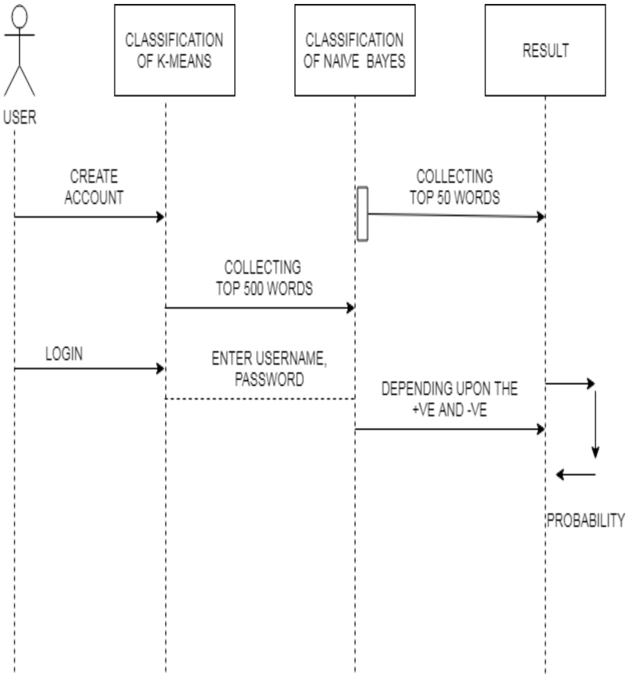

Figure 1: Sequence diagram

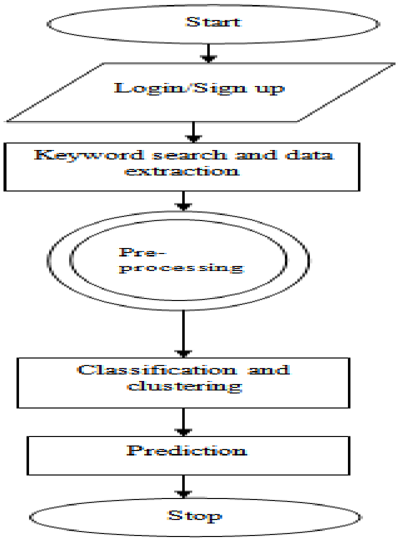
Start
Login/Sign up
Keyword search and data extraction
Pre-processing
Classification and clustering
Prediction
Stop

Figure 2: System Architecture of the project

IV. RESULT

The result of our project shows the polarity of a particular brand sentence based on the words given to the algorithm as training data. It shows the negativity or positivity in terms of probabilistic calculations.

V. CONCLUSION AND FUTURE WORKS

In this paper, a vector based anomaly detection model is proposed that accounts for efficient detection of anomalies in online social networks. The proposed model uses support vector machine and detect anomaly action as soon as performed (Live).Study results indicated that SVM based anomaly detection algorithm is efficient in identifying anomaly actions and the visualization is useful for analysts to discover insights and comprehend the model. The proposed approach uses simple algorithm even for identifying complex anomalies and mixed posts.

In the future, we will further investigate anomaly detection models for Twitter conversational threads and improve the current algorithm to allow a faster analysis. In addition to anomaly detection, it is interesting to integrate other content features (e.g., topics and semantic information) to the current system. Results suggest that the proposed model proves to be efficient and more accurate incomparison with the existing approaches over various parameters namely, anomaly filtering rate,accuracy in anomaly detection, convergence value, and approach failures.The outcome of text-based decision making contributes to the development of any company products. Analysation of thousands of twitter based tweets provide a strong means for sentiment analysis, through which we come to know whether it has a negative or a positive impact in the society.So,it is beneficial for both customer as well as the company also.

In this paper we have limited our work only to English language.So, future work would be to implement in other languages also(French,German,etc.).Also we can use image processing with machine learning to provide the same result by image classification by improving the overall accuracy.

ACKNOWLEDGMENT

Firstly, we express our sincere thanks to our guide Mrs. Mangala C N, Assoc Professor, Department of CSE, EWIT and Dr. Arun Biradar, Head, Department of computer science and engineering for their moral support. We express our sincere gratitude to our principal Dr. K Chennakeshavalu for his constant support and encouragement, we also thank all the faculties of East West Institute of Technology for their co-operation and support.

REFERENCES

[1] "Application of text classification and clustering of Twitter data for business analytics", Alrence Santiago Halibas ; Abubucker Samsudeen Shaffi ; Mohamed Abdul Kader Varusai Mohamed,2018 Majan International Conference (MIC), 2018.

[2] "Text Analytics Market by Component (Software, Services), Application (Customer Experience Management, Marketing Management, Governance, Risk and Compliance Management), Deployment Model, Organization Size, Industry Vertical, Region - Global Forecast to 20," 2017.

[3] H. S. Yaram, "Machine learning algorithms for document clustering and fraud detection," in Proceedings of the 2016 International Conference on Data Science and Engineering, ICDSE 2016, 2017.

[4] L. P. Khobragade and V. Jethani, "Sentiment Analysis of Movie Review," Int. J. Adv. Res. Comput. Sci., vol. 8, no. 5, p. 19411948, 2017.

[5] R. A. Jain and P. Dandannavar, "text analytics framework using apache spark and combination of lexical and maching learning techniques," Int. J. Bus. Anal. Intell., vol. 5, no. 1, pp. 36–42, 2017.

[6]S.M. Allahyari et al., "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques," arXiv Prepr. arXiv, vol. 1707, no. 2919, pp. 1–13, 2017.

[7] V. K. S. Rawat, "Comparative Analysis of Data Mining Techniques, Tools and Maching Learning Algorithms for Efficient Data Analytics," JOSR J. Comput. Eng., vol. 19, no. 4, pp. 56–60, 2017.

[8] P. Gonçalves and M. Araújo, "Comparing and combining sentiment analysis methods," Proc. first ACM Conf. Online Soc. networks. ACM, pp. 27–38, 2013.

[9] W. H. Kaur and V. Mangat, "Dictionary based Sentiment Analysis of Hinglish text," Int. J. Adv. Res. Comput. Sci., vol. 8, no. 5, pp. 816–822, 2017.

[10] C. F. N. Ribeiro, M. Araújo, P. Gonçalves, M. André Gonçalves, and F. Benevenuto, "SentiBench - a benchmark comparison of state-of-the- practice sentiment analysis methods," EPJ Data Sci., vol. 5, no. 1, 2016.

[11] D. A. Moreno and T. Redondo, "Text Analytics: the convergence of Big Data and Artificial Intelligence," Int. J. Interact. Multimed. Artif. Intell., vol. 3, no. 6, p. 57, 2016.

[12] E. V. A. Kharde and S. S. Sonawane, "Sentiment Analysis of Twitter Data: A Survey of Techniques," Int. J. Comput. Appl., vol. 139, no. 11, pp. 975–8887, 2016.

[13] G. L. Ziora, "The sentiment analysis as a tool of business analytics in contemporary organizations," Stud. Ekon., pp. 234–241, 2016.

[14] K. O. Muller, I. Junglas, S. Debortoli, and J. Von Brocke, "Using Text Analytics to Derive Customer Service Management Benefits from Unstructured Data," MIS Q. Exec., vol. 15, no. 4, pp. 64–73, 2016.

[15] T. A. Trevino, "Introduction to K-means Clustering," Datascience.com. 2016.

[16] X. M. H. Peetz, M. De Rijke, and R. Kaptein, "Estimating Reputation Polarity on Microblog Posts," Inf. Process. Manag., vol. 52, no. 2, pp. 193–216, 2016.

[17] I. N. Yussupova, M. Boyko, and D. Bogdanova, "A Decision Support Approach based on Sentiment Analysis Combined with Data Mining for Customer Satisfaction Research," Int. J. Adv. Intell. Syst., vol. 1&2, 2015.

[18] J. S. K. Markham, M. Kowolenko, and T. L. Michaelis, "Unstructured Text Analytics to Support New Product Development Decisions," Res. Technol. Manag., vol. 58, no. 2, pp. 30–39, 2015..

[19] F. W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams Eng. J., vol. 5, no. 4, pp. 1093– 1113, 2014.

[20] U. M. Hofmann; and R. Klinkenberg;, "RapidMiner: Data Mining Use Cases and Business Analytics Applications," Zhurnal Eksp. i Teor. Fiz., 2013.

Authors Profile

Ms. Kumari Ashmita Sinha is pursuing her 8 semester B.E in Computer Science & Engineering at East West Institute of Technology, Bengaluru, India. Her area of interest includes Machine Learning.

Ms. Maanasa VKP is pursuing her 8 semester B.E in Computer Science & Engineering at East West Institute of Technology, Bengaluru, India. Her area of interest includes Machine Learning and Web Technology.

Ms. Nikhitha SP is pursuing her 8 semester B.E in Computer Science & Engineering at East West Institute of Technology, Bengaluru, India. Her area of interest includes Machine Learning and Big data.

Ms. Ramya S is pursuing her 8 semester B.E in Computer Science & Engineering at East West Institute of Technology, Bengaluru, India. Her area of interest includes Machine Learning and Python.

Mrs. Mangala C N received the B.E degree in Computer Science and Engineering from NCET, Bengaluru, VTU in2006 and got M.Tech degree in Computer Science from RVCE, Bengaluru, India. She is currently working as Associate Professor in the Department of CSE, EWIT, and pursuing PhD in DSCE, Bengaluru, India. Her area of interest includes Image Processing, Data Mining and Big Data.