ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 1, 2019

OPINION MINING ON TWITTER

M.V.B.T. Santhi,

Associate Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, India Email: santhi_ist@kluniversity.in

Abstract –

This study focuses on the analysis of tweets, which are textual exchanges among users on the social media platform Twitter, serving as a platform for expressing opinions on various subjects. With 321 million Twitter users and approximately 6000 tweets generated per second, sentiment analysis becomes crucial in understanding the emotional context of these messages. Sentiment analysis, a facet of natural language processing, is employed to determine the emotional tone, categorizing content as positive or negative. The unstructured text for sentiment analysis is sourced from diverse online platforms like blogs, reviews, and comments. Through automatic, hybrid, or rule-based cleaning processes, the text is prepared for the subsequent stage. The evaluation of emotional content is conducted using algorithms derived from artificial intelligence, data mining, and machine learning. This research integrates natural language processing and machine learning techniques for sentiment analysis, employing methods such as Document Term Matrix. The study utilizes supervised machine learning algorithms, including Random Forest, Poisson, Multivariable, Gaussian, Bernoulli, and Simple Naive Bayes, for text classification.

Key Words: Sentiment Analysis, Machine Learning algorithms, Text Classification, Term Document Matrix, Twitter**.**

1. INTRODUCTION

Sentiment analysis examines a text or speech's emotional content. As social media company sites get better thanks to client reviews, it is quite advantageous. Sentiment analysis makes it simple to determine a customer's feelings and categorize them as positive or negative. Understanding an organization's strengths and weaknesses is beneficial.

1.1. SENTIMENT ANALYSIS TYPES

1. Subjectivity/Objectivity Sentiment-Analysis: The text classified into 2 categories that are subjective or objective.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 1, 2019

2. Feature/ Aspect-Based Sentiment Analysis: It identifies different sentiments related to the text in different aspects of the document.

1.2 APPROACHING TYPES

1. Rule-based Approach: This method ascertains the text's sentiment by using a thumb rule. It analyzes the sentiment using language techniques. It tends to have relatively low generalization but can perform well in a small domain.
2. Automatic Approach: This category includes ML and lexicon approaches. ML classifies texts using both supervised and unsupervised methods.
3. Lexicon-based approaches, on the other hand, employ both unsupervised methods and sentiment lexicons—lists of words that convey people's opinions.
4. Hybrid Approach: It combines Lexicon and ML techniques. The key component of the hybrid method is lexicons.

In this case, we compared the accuracy of Poisson Naïve Bayes with other supervised algorithms for sentiment analysis. The Poisson Naïve Bayes model accounts for random events. We were encouraged to apply this technique for text classification because Poisson naïve bayes is not employed for text classification.

Sentiment analysis was conducted using Natural Language methods as well. By leveraging natural language to connect machines and people, it is used for interaction. NLP techniques analyze the provided text, comprehend its relationships, and investigate its synergy. NLP was utilized for both text classification and data cleaning.

2. LITERATURE SURVEY

In the study by Anuja Jain et al. [1], a comprehensive approach to sentiment analysis was presented, utilizing Apache Spark for text analysis. Their framework incorporated machine learning algorithms, specifically Naïve Bayes and Decision Trees, with notable success demonstrated by the effectiveness of Decision Trees. Ali Hasan et al. [2] focused on a Twitter dataset comparing political parties, employing SentiWordNet and WordNet to determine positive and negative scores. Maximum Entropy and Support Vector Machine methods were applied for sentiment analysis, with Support Vector Machine exhibiting the highest accuracy. Lei Zhang et al. [3] provided an overview of deep learning approaches for sentiment analysis, employing convolutional neural networks alongside recursive, recurrent, and autoencoder types. Bhumika Gupta et al. [4] utilized maximum entropy classifier, support vector

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 1, 2019

machines, and Bayesian-logistic regression for sentiment analysis, introducing a generalized technique based on Python.

In their study [5], Marium Nafees et al. employed support vector machines and logistic regression for sentiment analysis of product reviews, discussing performance at sentence, concept, and document levels. Apoorv Agarwal et al. [6] addressed two sentiment categorization tasks, developing Unigram and Tree kernel models for sentiment analysis, and providing a detailed feature analysis. Yulan et al. [7] introduced the Joint Sentiment Model, a probabilistic framework utilizing Latent Dirichlet Allocation to identify both topic and sentiment in text, applied to a dataset of movie reviews. Alexander Pak et al. [8] gathered a Twitter dataset using emoticons and employed n-grams and Support Vector Machine techniques for sentiment analysis.Tony Mullen et al. [9] used topic proximity, sentiment orientation, and PMI for sentiment analysis, incorporating support vector machines as a supervised machine learning approach. Songbo Tan et al. [10] utilized frequently cooccurring entropy and the Adaptive Naïve Bayes model for sentiment analysis, demonstrating its superiority over the Naïve Bayes Transfer classifier. Akshi Kumar et al. [11] discussed sentiment mining from text using dictionaries and corpora, providing a linear equation for determining overall text emotion. Swati Redhu et al. [12] offered an overview of machine learning methods for sentiment analysis, including KNN, SVM, and Naïve Bayes, applying text mining and classifiers to specific data subtasks.Oskar Ahlgren [13] explored VOSViewer and Keyword Analysis, creating maps based on keyword co-occurrence, and employed unsupervised Latent Dirichlet Allocation for sentiment analysis. Priyanka Tyagi et al. [14] discussed the Maximum Entropy method and Ensemble Classifier, along with dictionary- and corpus-based sentiment analysis techniques, applied to online reviews gathered from Twitter. Abhishek Kaushik et al. [15] utilized various supervised and unsupervised methods, including Latent Semantic Analysis and Case-based Reasoning, providing a comprehensive summary of sentiment analysis methods and resource.

3. METHODOLOGY

The process of extracting the raw text from the Twitter dataset is used to do sentiment analysis on data. Preparing the data is necessary in order to use it in the algorithms.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 1, 2019

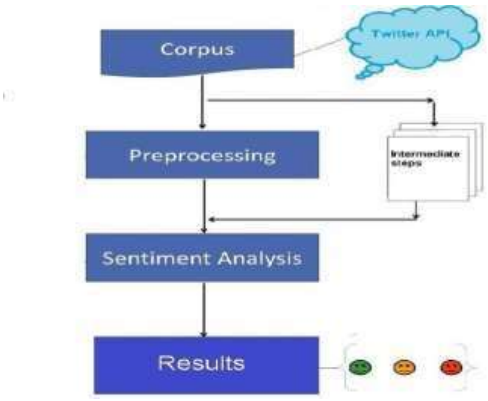

Figure 1 General Approach for Sentiment Analysis.

3.1 DATASET

A developer account must be created in order to access the Twitter API for tweet collecting. Alternatively, a Twitter dataset containing particular tweets may be used. One can obtain Twitter datasets through internet resources. For tweets that are downloaded from Kaggle, we have utilized the Apple dataset. There are 11 rows and 3887 columns in it.

3.2 PRE PROCESSING

Preprocessing the data is a crucial step since it prepares the raw data for cleaning. The processed data is utilized in subsequent procedures. For pre-processing, we have employed the Natural Language Processing approach. We used text mining techniques for pre-processing the dataset, removing stop words, URLs, numbers, and punctuations, and converting all characters to lowercase.

3.3 TERM DOCUMENT MATRIX

The category of Natural Language Processing includes Term Document Matrix. The text document is transformed into a two-dimensional matrix by it. In this matrix, terms are represented by rows, while documents are represented by columns. As a result, the total matrix shows the number of times each term appears in a specific matrix. Following this stage, a term-document matrix is used to complete the text classification.

3.4 CLASSIFIERS

Predicting the class of given data points is the process of classification. To determine which classifier has the best accuracy among the ones we employ, classifiers are used. Six supervised machine learning techniques are employed in this work.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper

 UGC CARE Listed (Group -I) Journal Volume 8, Issue 1, 2019

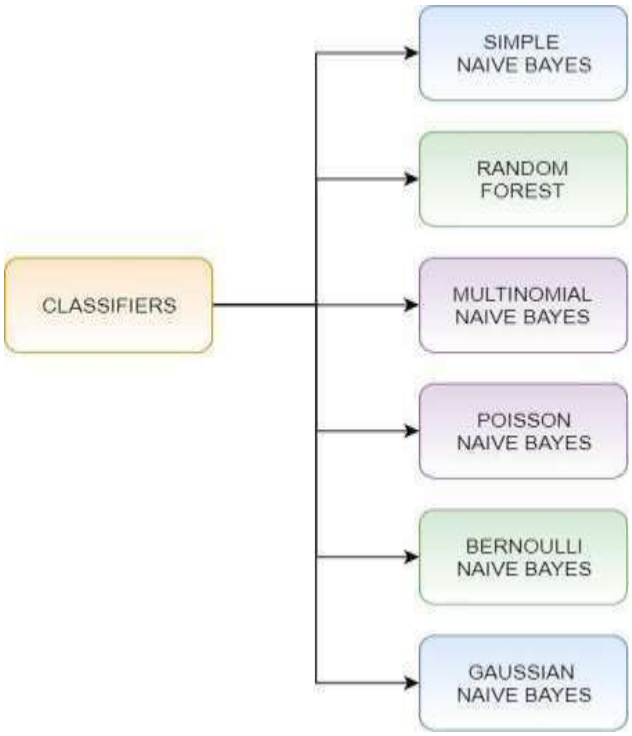

Figure 2 Classifiers used for analysing sentiments.

3.5 Naïve bayes

The Bayes theorem is necessary for Naïve Bayes to work. This classification method is predicated on the idea that predictors don't affect one another. It is simple to construct and helpful for large datasets.

3.6 Random Forest

There are several decision trees in this classifier. Every decision tree in this method offers an input classification. After compiling all of these, Random Forest selects the forecast with the highest number of votes and announces the outcome.

3.7 Multinomial Naïve bayes

Its primary function is text classification. It is referred to as a customized Naïve Bayes model. This category addresses word counts in documents and handles calculations inside them.

3.8 Poisson Naïve bayes

The purpose of this classifier is to simulate the random numbers of occurrences of a phenomenon at a given time. The Poisson Distribution is used to simulate phrase frequencies in documents where terms appear at random.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 1, 2019

3.9 Bernoulli Naïve bayes

Although it is mostly used to forecast Boolean variables, it is somewhat similar to the Multinomial NB type. This classifier forecasts if a given class variable will be true or false.

3.10 Gaussian Naïve bayes

For continuous data, this classifier is employed. Rather to discrete variables, continuous variables are used by the predictors.

3.11 WORD CLOUD

These are sometimes referred to as text clouds or tag clouds. It is a group of words rendered in various font sizes. A term is deemed significant if it occurs frequently in the paper. In the word cloud, that word is bolded and larger. As a result, the frequency of a word in the document determines its size in a word cloud. We generated a word cloud based on the term document matrix in this paper.

4. EXPERIMENTAL RESULTS

From the term-document matrix. A two-dimensional matrix is created which is used to represent repeated occurrences of the word in a given record. A bar plot was used to show the words which appeared frequently in the twitter_dataset.

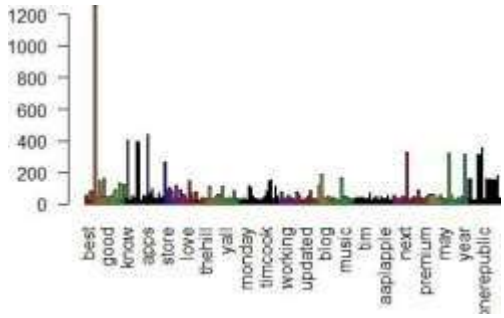

Figure 3 Bar Graph for term occurrences.

Six different Machine Learning algorithms are used for sentiment analysis. In which Poisson distribution Naïve Bayes and Random Forest performed well for the dataset we took. As Poisson Naïve Bayes generally not used for text classification, we did it on text classification and accuracy we got from this classifier is better compared to others.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper

 UGC CARE Listed (Group -I) Journal Volume 8, Issue 1, 2019

Figure 4 All classifiers collective results.

A word cloud is used to show the words from the twitter dataset. The size of the word in the word cloud is dependent on the word frequentness in the given data file.

Figure 5 Word Cloud of the data

A final bar graph is used to show the sentiments in the dataset. It represents eight different emotions associated with tweets in the dataset along with Positive and Negativesentiment.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 1, 2019

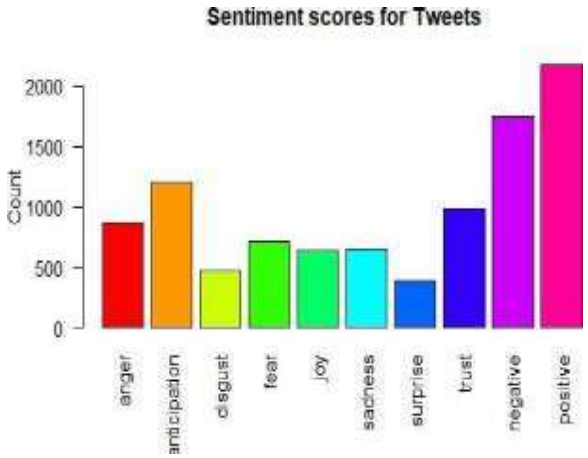

Figure 6 Bar graph representing sentiment scores of tweets**.**

5. CONCLUSION

This paper utilized the Apple dataset, comprising text-format tweets, and employed the Term Document Matrix in conjunction with various machine learning algorithms for text classification. Through extensive testing of classifiers, the Poisson distribution, Naïve Bayes, and Random Forest algorithms emerged as particularly effective, exhibiting the highest accuracy among supervised algorithms. Notably, the Poisson Naïve Bayes algorithm, less commonly applied in text classification, demonstrated superior performance compared to alternative algorithms, leading to its selection as the primary algorithm. Moving forward, potential avenues for future research involve incorporating audio and video data files, as well as datasets containing emoticons, from social media platforms beyond Twitter. Additionally, leveraging machine learning and advanced natural language techniques for sentiment analysis on diverse datasets from other social media platforms presents an exciting prospect for further investigation.

6.REFERENCES

[1] Anuja Jain and Padma Dandannavar. Application of Machine Learning Techniques to Sentiment Analysis, IEEE, 2017.

[2] Ali Hasan, Sana Moin , Ahmad Karim and Shahaboddin Shamshirband. Machine Learning-Based Sentiment Analysis for Twitter Accounts, MDPI, 2018.

[3] Lei Zhang, Shuai Wang and Bing Lui. Deep Learning for Sentiment Analysis, Cornell University, arXiv, 2018.

[4] Bhumika Gupta, Monika Negi .et.al. Study of Twitter Sentiment Analysis using Machine

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 1, 2019

Learning Algorithms on python, International Journal of Computer Applications Volume 165 – No.9, May 2017.

[5] Marium Nafees, Hafsa Dar, Salman Tiwana, Ikram Ullah Lali. Sentiment Analysis of Polarity in Product reviews in Social Media, ResearchGate, 2018.

[6] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow and Rebecca Passonneau. Sentiment Analysis of Twitter Data, Columbia University, 2017.

[7] Chenghua Lin and Yulan. Joint Sentiment/Topic Model for Sentiment Analysis, ACM, 2009.

[8] Alexander Pak, Patrick Paroubek. Twitter as a Corpus for Sentiment Analysis and Opinion Mining, crowdsourcingclass, 2017.

[9] Tony Mullen and Nigel Collier. Sentiment analysis using support vector machines with diverse information sources, aclweb, 2016.

[10] Songbo Tan, Hongbo Xu .et.al. Adapting Naive Bayes to Domain Adaptation for Sentiment Analysis, ResearchGate, 2009.

[11] Akshi Kumar and Teeja Mary Sebastian. Sentiment Analysis of Twitter Data, IJCSI International Journal of Computer Science Issues, Vol. 9, July 2012.

ISSN PRINT 2319 1775 Online 2320 7876

© 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 8, Issue 1, 2019

[12] Swati Redhu , Sangeet Srivastava, Barkha Bansal and Gaurav Gupta. Sentiment Analysis Using Text Mining, Science publishing group, 2018.Oskar Ahlgren. Research on Sentiment Analysis, Sentic, 2016.

[13] Priyanka Tyagi and Dr. R.C. Tripathi. A Review towards the Sentiment Analysis Techniques for the Analysis of Twitter Data, SSRN, 2019.

[14] Abhishek Kaushik, Anchal Kaushik and Sudhanshu Naithani. A Study on Sentiment Analysis: Methods and Tools, IJSR, 2014.

[15] Anila, M. & Pradeepini, G.2017, "Study of prediction algorithms for selecting appropriate classifier in machine learning". Journal of Advanced Research in Dynamical and Control Systems, vol. 9,no. Special Issue 18, pp. 257-268

[16] Muthukumaran, S., Suresh, P. & Amudhavel, J.2017, "Sentimental analysis on online

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 1, 2019

product reviews using LS-SVM method", Journal of Advanced Research in Dynamical and Control Systems, vol.9, no. Special Issue12, pp. 1342-1352.

[17] Mohiddin, S.K., Kumar, P.S., Sai, S.A.M., Santhi, M.V.B.T.Santhi ,2019, "Machine learning techniques to improve the results of student performance","International Journal of Innovative Technology and Exploring Engineering, Volume 8, Iaaue 5, 2019, Pages 590-594.

[18] Kodali, S., Dabbiru M. & Rao, B.T. 2018, "A survey of Data Mining techniques on information networks", International Journal of Engineering and Technology (UAE), vol.7,pp. 293- 300.

[19] Kousar Nikhath, A. & Subrahmanyam, K.2019,"Feature selection, optimization and clustering strategies of text documents", International Journal of Electrical and Computer Engineering, vol. 9, no.2, pp. 1313-1320.

[20] Krishna Mohan, G., Yoshitha, N., Lavanya, M.L.N. & Krishna Priya, A.2018, "Assessment and analysis of software reliability using machine learning techniques", International Journal of Engineering and Technology(UAE), vol.7, no. 2.32 Special Issue 32,pp. 201-205.

[21] Lakhmi Prasanna, P., RajeswaraRao, D.,Meghana, Y., Maithri, K. & Dhinesh, T. 2018, "Analysis of supervised classification techniques", International Journal of Engineering and Technology(UAE), vol. 7, no. 1.1, pp. 283-285.

[22] LaxmiNarasamma, V. & Sreedevi, M.2017, "A framework to analysis of tweets using multi-level tree algorithms", Journal of Advanced Research in Dynamical and Control Systems, vol.9, no. Special Issue 18, pp. 140-153.

[23] Narasinga Rao, M.R., Sajana, T., Bhavana, N., Sai Ram, M. & Nikhil Krishna, C.2018, "Prediction of chronic kidney disease using machine learning technique", Journal of Advanced