TWITTER OPINION MINING

M.V.B.T Santhi[1], Meghana. U[2], K. Dinesh[3]

[1,2,3]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Guntur, India.

---***---

Abstract - *Twitter is the online networking site, in it users interact with each other with texts known as Tweets. It is used to express people's emotions on a particular subject. There are 321 million active users and around 6000 tweets are tweeted every second on Twitter. Sentiment Analysis comes under Natural Language Processing. It is generally used to identify the sentiment of the text in the given document. It evaluates either the given text or speech is Positive or Negative. Unorganized text is gathered for sentiment analysis from various online sources such as blogs, reviews, comments, etc. The text is then cleaned and used for the further process using Automatic, Hybrid or Rule-based algorithms. Machine Learning, Data Mining, and Artificial Intelligence algorithms are used to analyze the emotion of the given text. In this paper, for Sentiment Analysis we used Natural Language techniques and Machine Learning together. We have used Machine Learning supervised algorithms namely Simple na�e bayes, Poisson na�e bayes, Multivariable na�e bayes, Gaussian na�e bayes, Bernoulli na�e bayes, Random Forest along with Document term Matrix for text classification*

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

1. INTRODUCTION

Sentiment-Analysis analyses emotion of text or speech. It is highly beneficial for Social Media business sites as they improve based on customer reviews. Sentiment Analysis easily identifies customers' emotions and classifies whether the sentiment is Positive or Negative. It helps organizations to know their Strong and Weak points.

1.1. SENTIMENT ANALYSIS TYPES

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

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 approach uses thumb rule to determine the sentiment of the text. It uses linguistic methods to analyze the sentiment.

2. It can have good performance with a narrow domain but tends to have a very poor generalization.

3. Automatic Approach: Lexicon and ML techniques comes under Automatic approach. ML uses supervised and unsupervised techniques for text classification. Whereas, Lexicon based approach uses unsupervised techniques and also uses sentiment lexicons which consist list of words that express people's opinions. 4. Hybrid Approach: It involves both ML and Lexicon methods together. Lexicons plays the major role in hybrid approach.

In this, we have used Poisson Na�e Bayes for Sentiment Analysis and compare its accuracy with other supervised Algorithms. Poisson Na�e Bayes is used to model for random occurrences. As Poisson na�e bayes not used for Text classification, this motivated us to use this algorithm for Text classification.

We also used Natural Language techniques for Sentiment Analysis. It's used to interact by connecting the computers and humans using the Natural-Language. NLP Techniques breaks down the given text, understands the relation between them and explore how they work together. We used NLP to clean the given data and also for text classification.

2. LITERATURE SURVEY

In [1] Anuja Jain .et.al gave a detailed approach of Sentiment Analysis. In their work, they proposed a text analysis framework using Apache spark. They have used Machine Learning algorithms Na�e Bayes and Decision Trees in their proposed framework. The results in the paper shown that Decision Trees had performed extremely well.

In the paper [2] Ali Hasan .et.al they extracted a twitter dataset about a comparison of two political parties from Twitter API. They used SentiWordNet and WordNet to find positive and negative scores of the mentioned data. They have used Support Vector Machine and Maximum Entropy algorithms for sentiment analysis. Support Vector Machine has shown the highest accuracy at the end.

From [3] Lei Zhang .et.al provided an Deep Learning overview and survived on sentiment-analysis based on the techniques of Deep Learning.

They have used Neural networks, Convolutional Neural Networks. Also Auto Encoder, Recursive and Recurrent Types.

In [4] Bhumika Gupta .et.al used ML algorithms like Bayesian-logistic regression, Support Vector Machine and also Maximum entropy classifier for sentiment analysis.

They also mentioned a python based generalized approach in their paper.

In their paper [5] Marium Nafees .et.al discussed Logistic Regression and Support Vector Machine for sentiment analysis of product reviews. They mentioned about performance of sentiment analysis in Sentence, Concept and Document levels. They applied sentiment analysis on Text classification and Emoticons classifications.

In paper [6] Apoorv Agarwal .et.al mentioned two sentiment classification tasks. One is about Positive and Negative sentiments, Another one is Positive, Negative and Neutral sentiments. They have proposed Unigram and Tree kernel models for the sentiment analysis task. They also presented the feature analysis of a hundred features.

In [7] Yulan .et.al proposed a probabilistic novel framework using Latent Dirichlet Allocation called Joint sentiment model that detects not only the sentiment but also the topic of the given text. It is an unsupervised model and they used this model on Movie reviews dataset.

From the paper [8] Alexander Pak .et.al used emoticons as a dataset that is collected from Twitter API for sentiment analysis. They used two types of emoticons Sad and Happy. They classified the data using Support Vector Machine and n-grams algorithms.

In their paper [9] Tony Mullen .et.al used sentiment orientation with PMI and Topic proximity techniques for sentiment analysis. Along with these techniques they used a Machine Learning supervised algorithm called support vector machine.

In [10] Songbo Tan .et.al performed sentiment analysis using the Adaptive Na�e Bayes model along with Frequently cooccurring entropy. The experimental results in their paper showing that this model provided better performance than Na�e Bayes Transfer cla ssifier.

In their paper [11] Akshi Kumar .et.al discussed about mining sentiment from the text. They mentioned Corpus and Dictionary methods for analysing the sentiments. They proposed a Linear equation through which they calculated the overall sentiment of the text.

In [12] Swati Redhu .et.al gave an overview of different Machine Learning algorithms KNN, Support Vector Machine(SVM) and na�e bayes for sentiment analysis. They have used Text Mining on the given data and then used classifiers on subtasks of the data for sentiment analysis.

In the paper [13] Oskar Ahlgren mentioned VOSViewer and an Analysis of Keyword. VOSViewer program creates maps based on the occurrence of two keywords together. He also used unsupervised algorithm Latent Dirichlet Allocation for sentiment analysis. In the results, they have compared both Keyword Analysis and Latent Dirichlet Allocation.

In [14] Priyanka Tyagi .et.al mentioned the Maximum Entropy algorithm and Ensemble Classifier from Machine Learning. Also, they mentioned corpus and dictionary-based methods for sentiment analysis. They worked all these methods on online reviews data collected from the Twitter website.

From the paper [15] Abhishek Kaushik .et.al mentioned that they have used Latent Semantic Analysis, Case-based Reasoning for sentiment analysis along with other supervised and unsupervised algorithms. In their paper, they also discussed an overview of sentiment analysis with the required techniques and tools.

3. METHODOLOGY

Sentiment Analysis of data is performed by extracting the raw text from the Twitter dataset. To use this data in the algorithms, the pre-processing of the data should be done.

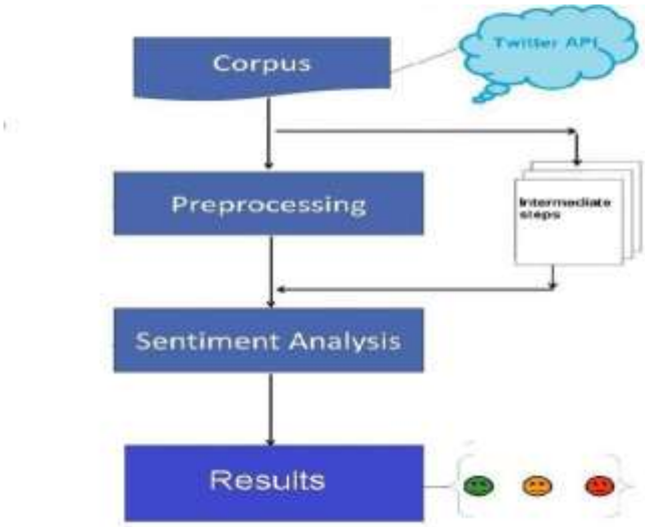

Fig.1 General Approach for Sentiment Analysis.

3.1 DATASET

Tweet collection can be done by gathering the tweet data from Twitter API by creating a developer account or by using a Twitter dataset that contains specific tweets. Twitter datasets can be downloaded from online sources. We have used the Apple dataset for tweets that are downloaded from Kaggle. It contains 3887 columns and 11 rows.

3.2 PRE PROCESSING

Preprocessing of the data is an important step as it makes raw data to be cleaned. This cleaned data is used for further process. We have used Natural Language Processing technique for pre-processing. In the dataset, we took, the pre-processing by using text mining technique removed Punctuations, URL's, Numbers, Stop words and converting letters into lower case.

3.3 TERM DOCUMENT MATRIX

Term Document Matrix comes under Natural Language Processing. It converts the text document into a two-dimensional matrix. The rows in this matrix represent the terms in a document and columns represent the documents. Therefore, the overall matrix represents how many occurrences of the terms are present in a particular matrix. After this step, the text classification is done with a termdocument matrix.

3.4 CLASSIFIERS

Classification is a process of predicting the class of data points that are given. Classifiers are used to find the highest accuracy classifier among the ones we use. In this paper, we have used Six Machine Learning supervised algorithms.

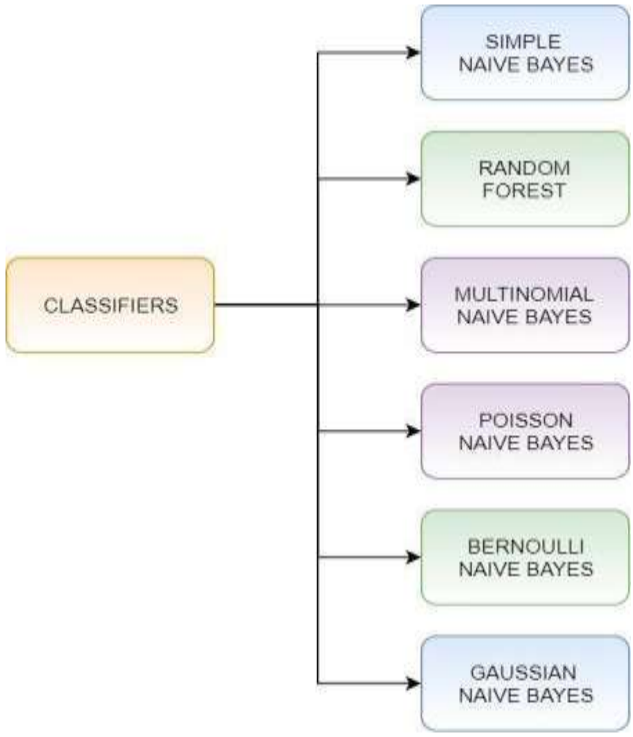

Fig.2 Classifiers used for analysing sentiments.

3.5 Na�e -bayes

Na�e -bayes depends on the Bayes theorem. This is the classification technique that assumes that predictors are independent of each other. It is useful for huge datasets and also easy to build.

3.6 Random Forest

This classifier involves many decision trees. In this algorithm, each decision tree provides a classification for input. Random Forest collects all these and chooses the most voted prediction and declares that as a result.

3.7 Multinomial na�e -bayes

It is mainly used for classification of the text. It's known as specialized version of Na�e Bayes. This is the classification which deals with word countsin a document and also works on calculations in it.

3.8 Poisson na�e -bayes

This classifier is used to model random occurrences numbers for a phenomenon at a specific time. If a term randomly occurs in a document then Poisson Distribution is used to model the term frequencies in the document.

3.9 Bernoulli na�e -bayes

It's quite similar to Multinomial NB type however it is generally used to predict the Boolean variables. This classifier predicts whether the given class variable subjects to yes or no.

3.10 Gaussian Na�e Bayes

This classifier is used for continuous values. The predictors work for continuous variables instead of discrete variables.

3.11 WORD CLOUD

These are also known as Tag clouds or Text clouds. It is a collection of words depicted in different sizes. If a word in a document appears frequently then it is considered as important. That word appears bigger and bolder in the word cloud. Therefore, the size of the word in a word cloud depends on the frequency of that word in the document. In this paper, we created a word cloud depending on the term document matrix.

4. SENTIMENT ANALYSIS WITH R

4.1 R language

R programming language is useful for statistics and graphics. R libraries are usually written in R, C, C++, and Fortran. Machine Learning algorithms and Data analysis are generally used in R. R is an easy and clear language. R libraries are mainly made by keeping data science in the mind.

4.2 Packages for Sentiment Analysis

In this paper, we are using R language for Sentiment Analysis because there are many R packages which make sentiment analysis task easier and R is easy to use. The packages we used are tm package (text mining), caret package for accuracy, word cloud package to create a word cloud, ggplot2 package for plots, syuzhe, dyplr and lubridate package for sentiments. We used nrc_sentiments command to get sentiments (Positive and Negative) along with eight other emotions for the given dataset.

5. 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.

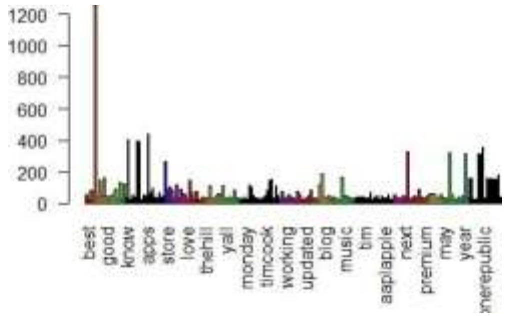

Fig.3 Bar Graph for term occurrences.

Six different Machine Learning algorithms are used for sentiment analysis. In which Poisson distribution Na�e Bayes and Random Forest performed well for the dataset we took. As Poisson Na�e 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.

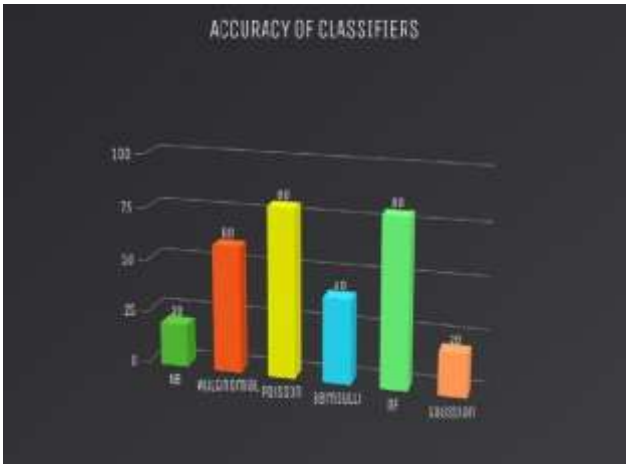

Fig.4 All classifiers collective results.

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.

aapl
onerepublic
iphone
future
one
computers
studio
need
now
outlet
batteries
store

Fig.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 Negative sentiment.

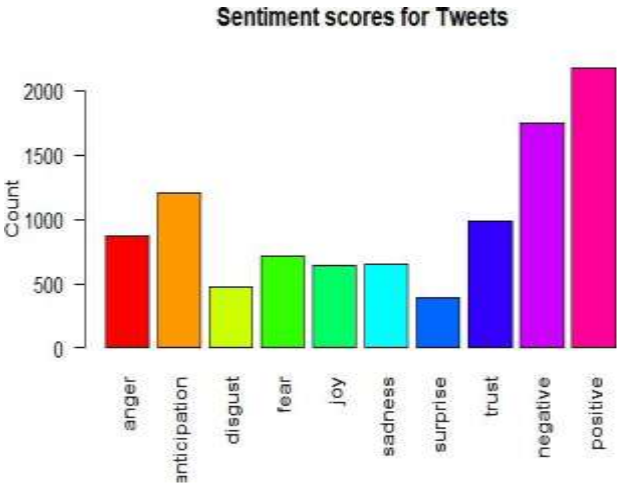

Fig.6 Bar graph representing sentiment scores of tweets.

6. CONCLUSIONS

In this paper, we collected tweets from the Apple dataset. This dataset contains tweets presented in the format of the text. We have used well known Machine Learning algorithms along with the Term document Matrix for text classification. After working out with different classifiers Poisson distribution Na�e Bayes and Random Forest algorithms are considered as efficient because they got maximum accuracy than other supervised algorithms. We considered Poisson Na�e Bayes as the main algorithm as it is not regularly used for text classification and compared it with other algorithms. In the sentiment analysis accuracy result, Poisson Na�e Bayes performed extremely well compared to other Na�e Bayes algorithms and equally with the Random Forest algorithm.

The future work can be done by using datasets with emoticons, audio, and video data files. Extracting the data from other social sites other than twitter and performing sentiment analysis on those by applying more Natural Language techniques and Machine Learning algorithms.

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 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.

[12] Swati Redhu , Sangeet Srivastava, Barkha Bansal and Gaurav Gupta. Sentiment Analysis Using Text Mining, Science publishing group, 2018.

[13] Oskar Ahlgren. Research on Sentiment Analysis, Sentic, 2016.

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

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

[16] 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

[17] Muthukumaran, S., Suresh, P. & Amudhavel, J.2017, "Sentimental analysis on online product reviews using LS-SVM method", Journal of Advanced Research in Dynamical and Control Systems, vol.9, no. Special Issue 12, pp. 1342-1352.

[18] Mohiddin, S.K., Kumar, P.S., Sai, S.A.M., Santhi, M.V.B.T. 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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

BIOGRAPHIES

M.V.B.T Santhi, Associate Professor of Computer Science and Engineering Department at KL University. Received B.Tech degree in Computer Science and Engineering from Jawaharlal Nehru Technological University in 2003, M.Tech degree in Computer Science from Archarya Nagarjuna University in 2010. Interested in Data Warehousing, Design and Analysis of Algorithms, DBMS, Big Data and Business Intelligence.

Meghana Uppaluri, Final year undergraduate student from Koneru Lakshmaiah Education foundation. Pursuing Computer Science and Engineering with a specialization in Computational Intelligence. Interested in the field of Machine Learning.

Korrapati Dinesh, Final year undergraduate student from Koneru Lakshmaiah Education foundation. Pursuing Computer Science and Engineering with a specialization in Computational Intelligence. Interested in the field of Machine Learning.