Sentiment analysis of twitter data using Machine learning algorithms

[1] **Karlapudi Naga Manasa**
[1]M.tech, Department of Computer Science,
Andhra University,Visakhapatnam, AP, India.

ABSTRACT: The rapid increase in usage of Technology has changed the way of expressing people's opinions, views and Sentiments about specific product, services, people and more, by using social media services such as Facebook, Instagram and Twitter. Due to this is massive amount of data gets generated. To find insights from this Data generated and make certain decision we implement web application that collects twitter data and shows it indifferent statistical forms. The main objective of the work presented with in this paper was to design and implement twitter data analysis and visualization in Python platform. Our primary approach was to focus on real-time analysis rather than historic datasets. Twitter API allow for collecting the sentiments information in the form of either positive score, negative score or neutral. We show the application of sentimental analysis and how to connect toTwitter and run sentimental analysis queries. We run experiments on different queries from politics to humanity and show the interesting results. We realized that the neutral sentiment for tweets are significantly high which clearly shows the limitations of the current works. this study focuses mainly on sentiment analysis of twitter data which is helpful to analyze the information in the tweets where opinions are highly unstructured, heterogeneous and are either positive or negative, or neutral in some cases. In this paper, we provide a survey and a comparative analyses of existing techniques for opinion mining like machine learning and lexicon-based approaches, together with evaluation metrics. Using various machine learning algorithms like Naive Bayes, XGBoost Classifier and Support Vector Machine, we provide research on twitter data streams. We have also discussed general challenges and applications of Sentiment Analysis on Twitter.

KEYWORDS:Twitter, Sentiment analysis (SA), Opinion mining, Machine learning, Naive Bayes (NB), XGBoost Classifier, Support Vector Machine (SVM).

I. **INTRODUCTION**

Nowadays, the age of Internet has changed the way people express their views, opinions. It is now mainly done through blog posts, online forums, product review websites, social media, etc. Nowadays, millions of people are using social network sites like Facebook, Twitter, Google Plus, etc. to express their emotions, opinion and share views about their daily lives. Through the online communities, we get an interactive media where consumers inform and influence others through forums. Social media is generating a large volume of sentiment rich data in the form of tweets, status updates, blog posts, comments, reviews, etc. Moreover, social media provides an opportunity for businesses by giving a platform to connect with their customers for advertising. People mostly depend upon user generated content over online to a great extent for decision making. For e.g. if someone wants to buy a product or wants to use any service, then they firstly look up its reviews online, discuss about it on social media before taking a decision. The amount of content generated by users is too vast for a normal user to analyze. So there is a need to automate this, various sentiment analysis techniques are widely used [1][2].

Sentiment analysis (SA) tells user whether the information about the product is satisfactory or not before they buy it. Marketers and firms use this analysis data to understand about their products or services in such a way that it can be offered as per the user's requirements.Textual Information retrieval techniques mainly focus on processing, searching or analyzing the factual data present. Facts have an objective component but [3], there are some other textual contents which express subjective characteristics. These contents are mainly opinions, sentiments, appraisals, attitudes, and emotions, which form the core of Sentiment Analysis (SA). It offers many challenging opportunities to develop new applications, mainly due to the huge growth of available information on online sources like blogs and social networks. For example, recommendations of items proposed by a recommendation system can be predicted by taking into account considerations such as positive or negative opinions about those items by making use of SA.

Opinion and sentimental mining is an important research areas because due to the huge number of daily posts on social networks, extracting people's opinion is a challenging task. About 90 percent of today's data has been provided during the last two years and getting insight into this large scale data is not trivial [12, 11].Sentimental analysis has many applications for different domains for example in businesses to get feedbacks for products by which companies can learn users' feedback and reviews on social medias. Opinion and sentimental mining has been well studied in this reference and all different approaches and research fields have been discussed[10]. There are also some works have been done on Facebook [10-12] sentimental analysis however in this paper we mostly focus on the Twitter sentimental analysis. For a larger texts one solution could be understand the text, summarize it and give weight to it whether it is positive, negative or neutral. Two fundamental approaches to extract text summarization are an extractive and abstractive method. In the extractive method, words and word phrases are extracted from the original text to generate a summary. In an abstractive method, tries to learn an internal language representation and then generates summary that is more similar to the summary done by human.

In this paper, we will discuss social network analysis and the importance of it, then we discuss Twitter as a rich resource for sentimental analysis. In the following sections, we show the high-level abstract of our implementation. We will show some queries on different topicsand show the polarity oftweets [8][5].

2. SENTIMENT ANALYSIS

Sentiment analysis can be defined as a process that automates mining of attitudes, opinions, views and emotions from text, speech, tweets and database sources through Natural Language Processing (NLP). Sentiment analysis involves classifying opinions in text into categories like "positive" or "negative" or "neutral". It's also referred as subjectivity analysis, opinion mining, and appraisal extraction.

The words opinion, sentiment, view and belief are used interchangeably but there are differences between them.

Opinion: A conclusion open to dispute (because different experts have different opinions)

View: subjective opinion

Belief: deliberate acceptance and intellectual assent

Sentiment: opinion representing one's feelings

An example for terminologies for Sentiment Analysis is as given below,

<SENTENCE> = the story of the movie was weak and boring

<OPINION HOLDER> =<author>

<OBJECT> = <movie>

<FEATURE> = <story>

<OPINION >= <weak><boring>

<POLARITY> = <negative>

Sentiment Analysis is a term that include many tasks such as sentiment extraction, sentiment classification, and subjectivity classification, summarization of opinions or opinion spam detection, among others. It aims to analyze people's sentiments, attitudes, opinions emotions, etc. towards elements such as, products, individuals, topics, organizations, and services.

Mathematically we can represent an opinion as a quintuple (o, f, so, h, t), where o =object; f =feature of the object o; so=orientation or polarity of the opinion on feature f of object o; h=opinion holder; t =time when the opinion is expressed.

Object: An entity which can be a, person, event, product, organization, or topic

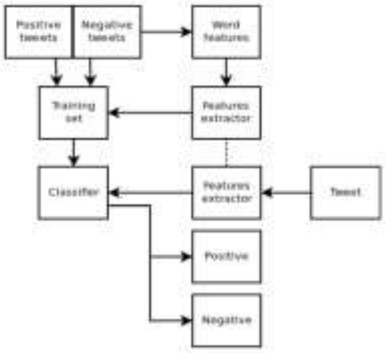
Feature: An attribute (or a part) of the object with respect to which evaluation is made.

Opinion orientation or polarity: The orientation of an opinion on a feature f represent whether the opinion is positive, negative or neutral .

Opinion holder: The holder of an opinion is the person or organization or an entity that expresses the opinion .

In recent years a lot of work has been done in the field of "Sentiment Analysis on Twitter" by number of researchers. In its early stage it was intended for binary classification which assigns opinions or reviews to bipolar classes such as positive or negative only.

Social networks is a rich platform to learn about people's opinion and sentiment regarding different topics as they can communicate and share their opinion actively on social medias including Facebook and Twitter. There are different opinion-oriented information gathering systems which aim to extract people's opinion regarding different topics. The sentiment-aware systems these days have many applications from business to social sciences. Since social networks, especially Twitter, contains small texts and people may use different words and abbreviations which are difficult to extract their sentiment by current Natural Language processing systems easily, therefore some researchers have used deep learning and machine learning techniques to extract and mine the polarity of the text.



Sentiment Analysis Architecture

2.1 Pre-processing of the datasets

A tweet contains a lot of opinions about the data which are expressed in different ways by different users .The twitter dataset used in this survey work is already labeled into two classes viz. negative and positive polarity and thus the sentiment analysis of the data becomes easy to observe the effect of various features. The raw data having polarity is highly susceptible to inconsistency and redundancy. Preprocessing of tweet include following points,

1. Remove all URLs (e.g. www.xyz.com), hash tags (e.g. #topic), targets (@username)
2. Correct the spellings; sequence of repeated characters is to be handled
3. Replace all the emoticons with their sentiment.
4. Remove all punctuations ,symbols, numbers
5. Remove Stop Words
6. Expand Acronyms(we can use a acronym dictionary)
7. Remove Non-English Tweets

2.2 Feature Extraction

The preprocessed dataset has many distinctive properties. In the feature extraction method, we extract the aspects from the processed dataset. Later this aspect are used to compute the positive and negative polarity in a sentence which is useful for determining the opinion of the individuals using models like unigram, bigram [18].

Machine learning techniques require representing the key features of text or documents for processing. These key features are considered as feature vectors which are used for the classification task.Some examples features that have been reported in literature are:

1. Words and Their Frequencies:

Unigrams, bigrams and n-gram models with their frequency counts are considered as features. There has been more research on using word presence rather than frequencies to better describe this feature. Panget al. [23] showed better results by using presence instead of frequencies.

2. Parts Of Speech Tags

Parts of speech like adjectives, adverbs and some groups of verbs and nouns are good indicators of subjectivity and sentiment. We can generate syntactic dependency patterns by parsing or dependency trees.

3. Opinion Words and Phrases

Apart from specific words, some phrases and idioms which convey sentiments can be used as features.e.g. cost someone an arm and leg.

4. Position of Terms

The position of a term with in a text can effect on how much the term makes difference in overall sentiment of the text.

5. Negation

Negation is an important but difficult feature to interpret. The presence of a negation usually changes the polarity of the opinion.

6. Syntax

Syntactic patterns like collocations are used as features tolearn subjectivity patterns by many of the researchers.

2.3 Training

Supervised learning is an important technique for solvingclassification problems. Training the classifier makes it easierfor future predictions for unknown data.

2.4 Classification

2.4.1 Naive Bayes

It is a probabilistic classifier and can learn the pattern ofexamining a set of documents that has been categorized [9]. Itcompares the contents with the list of words to classify thedocuments to their right category or class. Let d be the tweetand c* be a class that is assigned to d, where

$$C^{*} = \arg mac_{c} P_{NB}(c \mid d)$$

$$P_{NB}(c \mid d) = \frac{(P(c)) \sum_{i=1}^{m} p(f \mid c)^{n_i(d)}}{P(d)}$$

From the above equation,f‟ is a „feature‟, count of feature (fi)is denoted with ni(d) and is present in d which represents atweet. Here, m denotes no. of features.Parameters P(c) and P(f|c) are computed through maximumlikelihood estimates, and smoothing is utilized for unseenfeatures. To train and classify using Naïve Bayes MachineLearning technique ,we can use the Python NLTK library[09] .

2.4.2 XGBoostClassifier

In XGBoost Classifier, no assumptions are takenregarding the relationship in between the features extracted from dataset. This classifier always tries to maximize theentropy of the system by estimating the conditional distribution of the class label. XGBoost Classifier even handles overlap feature and is same as logistic regression method which finds the distribution over Classes. The conditional distribution is defined as MaxEntmakes no independence assumptions for its features, unlike[10]

Naive Bayes. The model is represented by the following:

$$P_{ME}(c \mid d, \lambda) = \frac{\exp[\sum_{i} \lambda_i f_i(c,d)]}{\sum_{c} \exp[\sum_{i} \lambda_i f_i(c,d)]}$$

Where c is the class, d is the tweet and λ_i is the weight vector. The weight vectors decide the importance of a feature in classification.

2.4.3 Support Vector Machine

Support vector machine analyzes the data, define the decision boundaries and uses the kernels for computation which are performed in input space [12]. The input data are two sets of vectors of size m each. Then every data which represented as a vector is classified into a class. Neatly we find a margin between the two classes that is far from any document. The distance defines the margin of the classifier, maximizing the margin reduces indecisive decisions. SVM also supports classification and regression which are useful for statistical learning theory and it also helps recognizing the factors precisely, that needs to be taken into account, to understand it successfully.

3. APPROACHES FOR SENTIMENTANALYSIS

There are mainly two techniques for sentiment analysis for the twitter data:

3.1 Machine Learning Approaches

Machine learning based approach uses classification technique to classify text into classes. There are mainly two types of machine learning techniques

3.1.1. Unsupervised learning:

It does not consist of a category and they do not provide with the correct targets at all and therefore rely on clustering.

3.1.2. Supervised learning:

It is based on labeled dataset and thus the labels are provided to the model during the process. These labeled dataset are trained

To get meaningful outputs when encountered during decision- making. The success of both this learning methods is mainly depends on the selection and extraction of the specific set of features used to detect sentiment. The machine learning approach applicable to sentiment analysis mainly belongs to supervised classification. In a machine learning techniques, two sets of data are needed:

1. Training Set
2. Test Set.

A number of machine learning techniques have been formulated to classify the tweets into classes. Machine learning techniques like Naive Bayes (NB), XGBoost Classifier(XC),and support vector machines (SVM) have achieved great success in sentiment analysis. Machine learning starts with collecting training dataset. Neatly we train a classifier on the training data. Once a supervised classification technique is selected, an important decision to make is to select feature. They can tell us how documents are represented.

The most commonly used features in sentiment classification are

 Term presence and their frequency

 Part of speech information

 Negations

 Opinion words and phrases

4. EVALUATION OF SENTIMENT CLASSIFICATION

The performance of sentiment classification can be evaluated by using four indexes calculated as the following equations:

$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN)$

$\text{Precision} = TP/(TP+FP)$

$\text{Recall} = TP/(TP+FN)$

$F1 = (2\times\text{Precision}\times\text{Recall})/ (\text{Precision}+\text{Recall})$

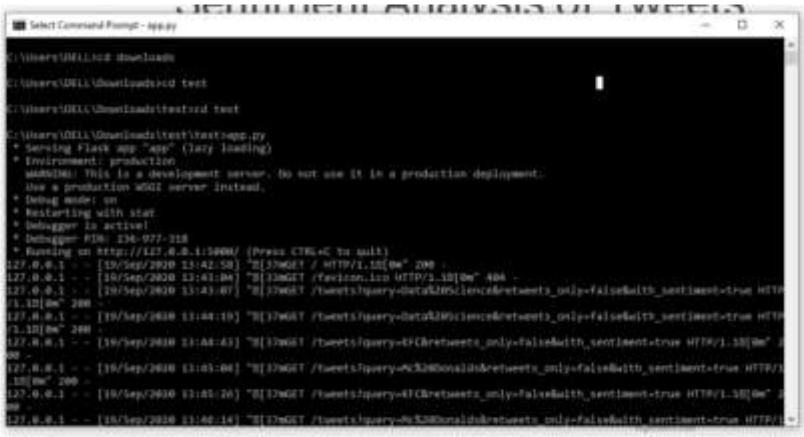
In which TP, FN, FP and TN refer respectively to the number of true positive instances, the number of false negative instances, the number of false positive instances and the number of true negative instances,

Confusion Matrix

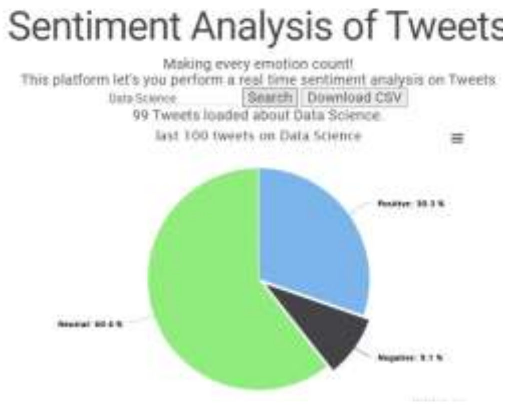
	Predicted Positives	Predicted Negatives
Actual Positive	TP	FN
Actual Negative	FP	TN

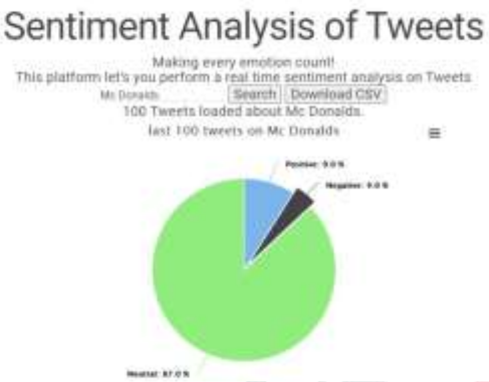
5. RESULTS AND DISCUSSION

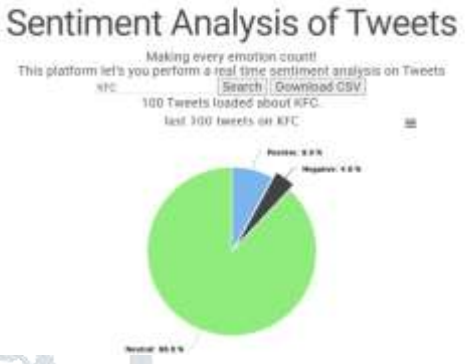
We used the twitter dataset publicly made available by Stanford University. Analyses was done on this labelled datasets using various feature extraction technique. We used the framework where the pre-processor is applied to the raw sentences which make it more appropriate to understand. Further, the different machine learning techniques trains the dataset with feature vectors and then the semantic analysis offers a large set of synonyms and similarity which provides the polarity of the content.



```
C:\Users\DELL>cd downloads

C:\Users\DELL\Downloads>cd test

C:\Users\DELL\Downloads\test>cd test

C:\Users\DELL\Downloads\test\test>app.py
 * Serving Flask app "app" (lazy loading)
 * Environment: production
   WARNING: This is a development server. Do not use it in a production deployment.
   Use a production WSGI server instead.
 * Debug mode: on
 * Restarting with stat
 * Debugger is active!
 * Debugger PIN: 236-977-318
 * Running on http://127.0.0.1:5000/ (Press CTRL+C to quit)
127.0.0.1 - - [19/Sep/2020 13:42:58] "←[37mGET / HTTP/1.1←[0m" 200 -
127.0.0.1 - - [19/Sep/2020 13:43:04] "←[33mGET /favicon.ico HTTP/1.1←[0m" 404 -
127.0.0.1 - - [19/Sep/2020 13:43:07] "←[37mGET /tweets?query=Data%20Science&retweets_only=false&with_sentiment=true HTTP
/1.1←[0m" 200 -
127.0.0.1 - - [19/Sep/2020 13:44:19] "←[37mGET /tweets?query=Data%20Science&retweets_only=false&with_sentiment=true HTTP
/1.1←[0m" 200 -
127.0.0.1 - - [19/Sep/2020 13:44:43] "←[37mGET /tweets?query=KFC&retweets_only=false&with_sentiment=true HTTP/1.1←[0m" 2
00 -
127.0.0.1 - - [19/Sep/2020 13:45:08] "←[37mGET /tweets?query=Mc%20Donalds&retweets_only=false&with_sentiment=true HTTP/1
.1←[0m" 200 -
127.0.0.1 - - [19/Sep/2020 13:45:26] "←[37mGET /tweets?query=KFC&retweets_only=false&with_sentiment=true HTTP/1.1←[0m" 2
00 -
127.0.0.1 - - [19/Sep/2020 13:46:14] "←[37mGET /tweets?query=Mc%20Donalds&retweets_only=false&with_sentiment=true HTTP/1
```







6. APPLICATIONS OF SENTIMENT ANALYSIS

Sentiment Analysis has many applications in various Fields.

1. Applications that use Reviews from Websites
2. Applications as a Sub-component Technology
3. Applications in Business Intelligence
4. Applications across Domains:
5. Applications in Smart Homes
6. Social Media Monitoring
7. Customer Service
8. Market Research
9. Brand Monitoring
10. Political Capaigns

CONCLUSION

In this paper, we provide a survey and comparative study of existing techniques for opinion mining including machine learning and lexicon-based approaches, together with cross domain and cross-lingual methods and some evaluation metrics. We discussed the importance of social network analysis and its applications in different areas. We focused on Twitter as and have implemented the python program to implement sentimental analysis. We showed the results on different daily topics. We realized that the neutral sentiments are significantly high which shows there is a need to improve Twitter sentiment analysis. Research results show that machine learning methods, such as SVM and naive Bayes have the highest accuracy and can be regarded as the baseline learning methods. We also studied the effects of various features on classifier.

FUTURE WORK

We can conclude that more the cleaner data, more accurate results can be obtained. Use of bigram model provides better sentiment accuracy as compared to other models. We can focus on the study of combining machine learning methods in order to improve the accuracy of sentiment classification and adaptive capacity to variety of domains and different languages.

REFERENCES

[1] Levy, M. (2016). Playing with Twitter Data. [Blog] R-bloggers. Available at: https://www.r-bloggers.com/playing-with-twitter-data/ [Accessed 7 Feb. 2018].

[2] Popularity Analysis for Saudi Telecom Companies Based on Twitter Data. (2013). Research Journal of Applied Sciences, Engineering and Technology. [online] Available at: http://maxwellsci.com/print/rjaset/v6-4676-4680.pdf [Accessed 1 Feb. 2018].

[3] Zhao, Y. (2016). Twitter Data Analysis with R – Text Mining and Social Network Analysis. [online] University of Canberra, p.40. Available at:https://paulvanderlaken.files.wordpress.com/2017/08/rdatamining-slides-twitter-analysis.pdf [Accessed 7 Feb. 2018].

[4] Alrubaiee, H., Qiu, R., Alomar, K. and Li, D. (2016). Sentiment Analysis of Arabic Tweets in e-Learning. Journal of Computer Science. [online] Available at: http://thescipub.com/PDF/jcssp.2016.553.563.pdf [Accessed 7 Feb. 2018].

[5] Qamar, A., Alsuhibany, S. and Ahmed, S. (2017). Sentiment Classification of Twitter Data Belonging to Saudi Arabian Telecommunication Companies. (IJACSA) International Journal of Advanced Computer Science and Applications, [online] 8. Available https://thesai.org/Downloads/Volume8No1/Paper_50-Sentiment_Classification_of_Twitter_Data_Belonging.pdf [Accessed 1 Feb. 2018].

[6] R. M. Duwairi and I.Qarqaz, “A framework for Arabic sentiment analysis using supervised classification” , Int. J. Data Mining, Modelling and Management, Vol. 8, No. 4, pp.369-381 , 2016.

[7] Hossam S. Ibrahim, Sherif M. Abdou, MervatGheith, “Sentiment Analysis For Modern Standard Arabic And Colloquial”, International Journal on Natural Language Computing (IJNLC), Vol. 4, No.2, pp. 95-109, April 2015.

[8] Assiri, A., Emam, A. and Al-Dossari, H. (2016). Saudi Twitter Corpus for Sentiment Analysis. International Journal of Computer and Information Engineering, [online] 10. Available at: http://waset.org/publications/10003483/saudi-twitter-corpus-for-sentiment-analysis [Accessed 1 Mar. 2018].

[9] L. Wasser and C. Farmer, "Sentiment Analysis of Colorado Flood Tweets in R", Earth Lab, 2018. [Online]. Available: https://earthdatascience.org/courses/earth-analytics/get-data-using-apis/sentiment-analysis-of-twitter-data-r/. [Accessed: 01- Mar-2018].

[10] D. Robinson, "Text analysis of Trump's tweets confirms he writes only the (angrier) Android half", Variance explained, 2016.

[11] _____"A Common Database Interface (DBI)", cran.r-R, 2003. [Online]. Available: https://cran.r-project.org/web/packages/DBI/vignettes/DBI-1.html. [Accessed: 25- Mar- 2018].

[12] V. Kharde and S. Sonawane, "Sentiment Analysis of Twitter Data: A Survey of Techniques", International Journal of Computer Applications, vol. 139, p. 11, 2016.