Sentimental Analysis on Twitter: Approaches and Techniques

[1]Akankasha, [2]Bhavna Arora
[1]M.Tech Student, [2] Assistant Professor
[1]Department of Computer Science & IT, Central University, Jammu,
[2]Department of Computer Science & IT, Central University, Jammu
[1]akankashaverma1000@gmail.com, [2]bhavna.aroramakin@gmail.com

ABSTRACT

Sentiment is a terminology which define an attitude, opinion, thought, or perception that indicates ones feeling. Sentiment analysis correspondingly called as opinion mining, facilitates the extraction of individual's sentiment towards certain elements. Now a days, social media applications like Twitter, Facebook etc. has an immense effect on individual lives as people post their thinking in form of posts on these applications. Researchers find Twitter to be the most commonly used social media application to post their opinions. Twitter is a microblogging sites in which a user send messages and those ongoing messages are called Tweets. But these sites carries various technical threats like noise, sparsity, non-standard vocabulary, multilingual content that is posted online. For tackling these challenges, the N-gram technique has been discussed which is used for feature extraction and Support Vector Machine (SVM) approach for classification for sentiment analysis. In this paper a brief introduction on Sentiment Analysis is given along with approaches and techniques. And a workflow on sentiment analysis technique also discussed.

Keywords: - Sentiment analysis, Opinion Mining, Social Media, N-gram technique, Support Vector Machine (SVM)

INTRODUCTION

Social Media (Facebook, Twitter, and Instagram) is a podium where individuals without obstruction can interact with each other and can share their opinions, personal experiences, reviews, ideas and messages. It is an electronic format (either mobile based or computer based). Over a year, Internet is playing an important role on once lives. It has changed perspective of a person on things. Commonly used social media application is Twitter. It is an online news and social networking site which enables users interact with one another, post comments, blogs and send messages and these ongoing messages are called as Tweets. Twitter is also called a microblogging site. Microblogs allows users to put down messages in real time about their opinions and review, analyse issues, and exaggerate products they use.

Twitter is universally known as a "what's-happening-right-now" tool. It is a platform which give power to its users and attract them to skylarking their thoughts and appreciate things happening in their lives in real-time.

Delving into the thought of millions of people can be useful and valuable source of data for analysis as well. With the help of Twitter, these manifest instantly obtainable in a data stream, that can be hewed by stream excavating methods. This possibly derive people's opinions, on an individual level and in combination or in both, with respect any matter or incident. At the official Twitter Chirp developer conference in April 2010, the company presented few statistics about its site and its users [1]. The significant mentions are:

- In April 2010, Twitter had 106 million sign up users and 180 million one of a kind steady visitors. The company uncovered that 300,000 new users were linking every day and that it received 600 million questions day by day by means of its search engine, and collection of 3 billion entreaties for each diurnal in view of its Application Program Interface(API).
- 37% of Twitter's vigorous users utilize their phone to send memoranda.

The underlying model for twitter input data is the data stream model. In this model, the data arrives fast. Algorithms for data mining on this data have been employed to predict data in real time, in conjunction with space and time. If, after a while the attributes of data changes data mining algorithms should have the capability to manage them [1].

Problems which can be solved with help of Twitter graph mining technique are stated below: –

1. Determining user influence and dynamics of admiration:- There are mainly three measures for determining user influence. These are in-degree, retweets and mentions.
2. Community discovery and formation**: -** It initiate societies that utilizes Hypertext Induced Topic Search (HITS), and the Clique Percolation Method. In this researchers analyzes the formation of links in Twitter by process of the directed closure process.
3. Social information diffusion**:-** In what way data sampling tactics affect encounter information diffusion is studied in this with the help of mining algorithm [2].

This paper is organized in five sections. After the introduction to social media application- Twitter in Section I. Section II describes the review of literature of the recent trends in the analysis of social media application and techniques applied. Sentiment Analysis, its approaches and techniques are discussed at length in section III. The work presented in this paper is a base line for research work based on sentimental analysis with reference to Twitter. The workflow on sentiment analysis has been conferred in section IV. Finally, the paper concludes and future work is presented in section V.

LITERATURE REVIEW

In [3], the author discusses the sentiment Analysis on Twitter using Streaming API. In this, the author has conversed that Twitter API is used to gather data from Twitter. It is an interactive

automatic method that prophesies the sentiment of the review/chirps that individuals post on social media using a tool called Hadoop.

In [4], the author proposed the N-gram graph model used in context of Sentiment Analysis over Social Media. The author has discussed various techniques like Term Vector Model, Character N-gram Model and N-Gram Model. All of the three techniques, N-gram has gained high priority because of the advantages: Firstly, it allows fuzzy and substring Matching. Secondly, it is a language-neutral Method.

The Sentiment analysis of individuals' sentiments with respect to top colleges in India has been discussed by the author in [5]. The author has done comparative analysis on Naïve Bayes and Support Vector Machine (SVM).Along with this a comparison has been made among 4 different kernels of Support Vector Machine (SVM): linear Classifier, polynomial classifier, Radial Basis Function (RBF), sigmoid.

In [6], the author has showed how Support Vector Machine (SVM) tools has been used for determining company ratings. The author has discussed various benefits of SVM like: It give precise and robust classification even if input data is non-monotonic and non- linear in nature also SVMs operates locally. Due to the following reasons, it is an operational tool for supplementing the data gained from classical linear for classification.

In [7], the author has discussed several sentiment analysis techniques like machine learning method, rule based approach and Lexicon based approach. Comparative search is also done on these methods. Machine learning method like SVM and Naïve Bayes offers best accuracy and considered as a best starting point for learning policies. In case of Lexicon approach, human characterized manuscript is mandatory and no need of learning strategy. And Rule based approach is collection of both techniques. It provides greater accuracy, involve less data but require skillful human effort.

SENTIMENT ANALYSIS

Sentiment analysis is a process that analyze, excerpt and excavate sentiment or opinion of users either from stanza, allocution or speech, tweets with the help of Natural Language Processing (NLP).It is a process which converts unstructured data into meaningful data. Sentiment is the study of meaning. Sentiment analysis classify opinions into polarity like positive, negative or neutral.

The figure below depicts the architecture of the sentimental Analysis [8]:

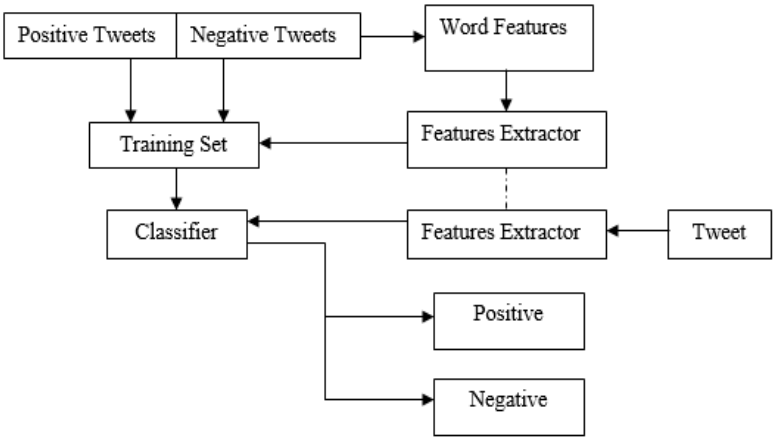

Fig.1 Sentiment Analysis Architecture

A. APPROACHES OF SENTIMENT ANALYSIS

There are two approaches for sentiment Analysis on Twitter Data as shown in the following figure [9]

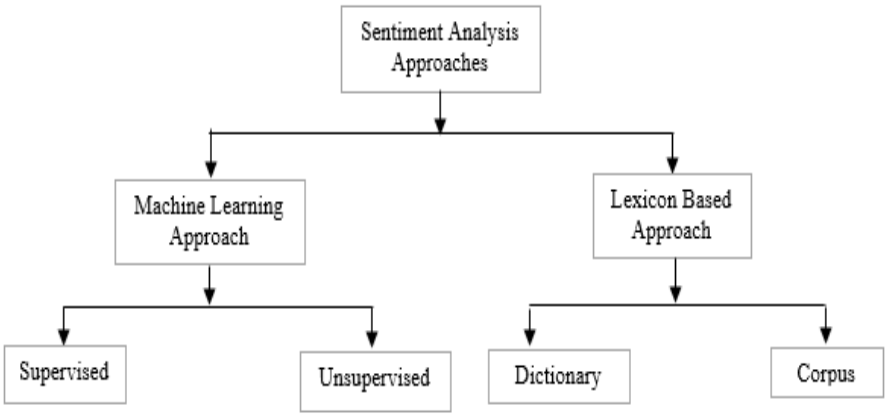

Fig.2. Approaches of Sentiment Analysis

I. **Machine Based Learning**:-It is a process in which a predefined dataset is given and an algorithm is trained on it before applying it on actual dataset. It uses classification technique to classify text into classes. It is classified into two categories.

a) **Unsupervised learning: -** It is a technique in which there is no predefined dataset is provided. There is no correct target and therefore rely on clustering.
b) **Supervised learning**: - In this technique a predefined data set is given. These trained dataset provide a genuine output when assembled at the time of judgement.

Two sets of data are used in machine learning technique [10]:

1. Training Set
2. Test Set.

Following machine learning techniques has been used for classifying tweets into categories like Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machines (SVM).

Features used in sentiment classification are given below:

- Words and its Frequencies
- Parts of Speech labeled
- Sentiment Words and Phrases
- Negation

II. **Lexicon Based Approach:-**This method concentrate on sentiment dictionary, a slant of words and phrases. It matches these dictionary opinions with the data to find the polarity of the words.

There are two sub classifications for this approach:-

a) **Dictionary-based Approach**:-In this method, a set of sentiment or opinion word is accepted, collected and illustrated manually. This happens by exploring the synonyms and antonyms of a dictionary. New word are added to the list and process goes on till no new word is found. Once the process is completed manual inquisition done to remove errors.

b) **Corpus-Based Approaches**:-This method deals with the context specific orientation. It basically work for the words which appear together like AND, OR, EITHER-OR, NETHER- NOR, etc. It also study real life languages which can be in form of stanza (structured text) or vernacular (audio files).

B. TECHNIQUES FOR SENTIMENT ANALYSIS

Two techniques being used for feature extraction and classification of twitter data. These techniques are discussed below.

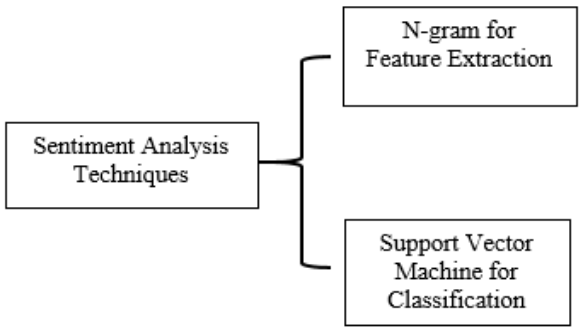

Fig.3. Sentiment Analysis Techniques

I. **N-Gram Technique:-** N-gram is a technique used for feature extraction over social media application. It is one of finest technique being used for extraction of useful content from an unstructured information. N-gram is simply adjoining order of word or letters of length n. In other words, N-gram is an arrangement of n objects from stated succession. It includes symbols, letters, words, phonemes etc. Size 1 N-gram is Unigram, size 2 is Bigram, and size 3 is Trigram and so on. These are collectively called Language Models (probability distribution over Sentences)

a) **Models used in Sentiment analysis technique :-**

- *Unigram Model*:-It is a collection of single token form a given string. The model is represented by following equation 1[11]

$$P(V)=P(v_1)\ P(v_2)\ (v_3)\ldots\ldots P(v_n)$$

Eqn 1. Unigram Model

For example, given statement "India is the Best" and apply all models.

Pictorial representation of Unigram Model

India	is	the	Best

Unigrams: - India, is, the, best

- *Bigram Model*: - It is a collection of two words from a given sentence. The model is represented by following equation 2[11]

$$P(V)=P(v_1)P(v_2/v_1)P(v_3/v_2)\ldots\ \ldots\ldots\ldots P(v_n/v_{n-1})$$

Eqn. 2.Bigram Model

Pictorial representation of Bigram Model

India is	is the	the Best

Bigram:- India is, is the, the best

- *Trigram Model*:- This language model collects three word of a string at once. This model is represented in equation 3[11]

$$P(V)=P(v_1)P(v_2/v_1)P(v_3/v_2,v_1)\ldots.P(v_n/v_{n-1}\ v_{N-2})$$

Eqn 3. Trigram Model

Pictorial representation of Trigram Model

India is the	Is the Best

Trigram Model: - India is the, is the Best

• *N-gram model*:-This module is a assortment of n item from a given sequence. It can be represented in equation 4[11]:-

$$\mathbf{P(V)= P(v_1)\ P(v_2/v_1)\ldots P(v_n/v_{n-1}\ v_{n-2}\ldots v_{n-N})}$$

Eqn.4. N-gram Model [11]

b) **Advantages of N-gram Technique:-**

1. Language independent.
2. Easy to use and effective.
3. Uses contextual information instead of having a plan of bag of word approach.
4. There is no need of text preprocessing

II. **Support Vector Machine (SVM)**:-It is a supervised machine learning technique. A powerful mechanism for 2-class classification and regression challenges. It is a non-probabilistic linear classifier. It plot the training data in multidimensional space and separate class with Hyperplane. The first step of Support vector machine is to analyze the data after this it distinguish the decision demarcation, at last it uses kernels for estimating performance on input zone. The input data is 2 sets of vectors about m size each. At that point each info which interpreted as a vector gets classifies as a class [12].

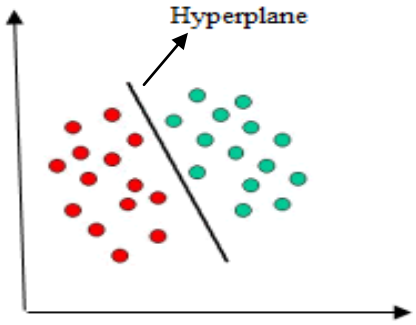

Fig 4. Linear Classifier [13]

a) **Parameters for Support Vector Machine**

SVM is divided into 4 parameters on the basis of classification and regression.

- **Methods for Classification**:- There are broadly two methods defined for classification as discussed below

1. **Type1 Classification***:* - It is also known as C-SVM. The C constraint controls the tradeoff amid decision boundaries. In addition it categorize training points appropriately. It ranges from 0 to infinite. Hence it is difficile to evaluation and customized [14] [15].
2. **Type2 Classification***:* - It is also known as nu-SVM, illustrated as ν .To overcome shortcoming of C-SVM this method was introduced. Nu runs between [0, 1] and epitomizes lower and upper bound on the number of instances that are support vector and that lie on the wrong side of the hyperplane [15].

- **Methods for regression**:-There are two methods defined for regression as discussed below.

1. **Type1 Regression***:*- It is also called as epsilon-SVM denoted as Ɛ. The significance of epsilon arbitrate the level of precision of the estimated function. It depend on the target values in the training set. Epsilon (Ɛ- value) distress model complexity [16].
2. **Type2 Regression:-**Another name of this method is nu-SVM denotes as ν. The parameter ν is used to determine fraction of support vector in the consequential model [16].

b) **Advantages of Support Vector Machine**

1. High Accuracy
2. Provides a unique solution
3. Fast evaluation.
4. Work when training examples contain error shows robustness[17].

WORK FLOW IN SENTIMENTAL ANALYSIS

The workflow of sentimental analysis is based on feature extraction and data classification. The primary step is features extraction and the extracted data or features are classified by classification techniques. The technique of N-gram has been used for the process of extraction of features from the site and the technique of Support Vector Machine (SVM) has been used for classifying the extracted features. Sentiment analysis technique steps elaborate below:-

1. **Input Data**: - Firstly, the data is collected and is provided as input. The given data is Twitter data The input data can be a real time data or data in excel sheet can be selected from twitty application
2. **Pre-processing**:-Secondly, the input data is preprocessed. Pre-processing of data embraces removing tokens, stop word, negation URLS and so on.
3. **Feature Extraction**: - The third step includes feature extraction. In this, N-gram algorithm is specified. The pre-processed data has been given as input to the n-gram module.

4.**Classification**: - Lastly, the classification approach has been applied on the extracted data for the sentiment analysis. For this purpose, Support Vector Machine classifier has been adopted.

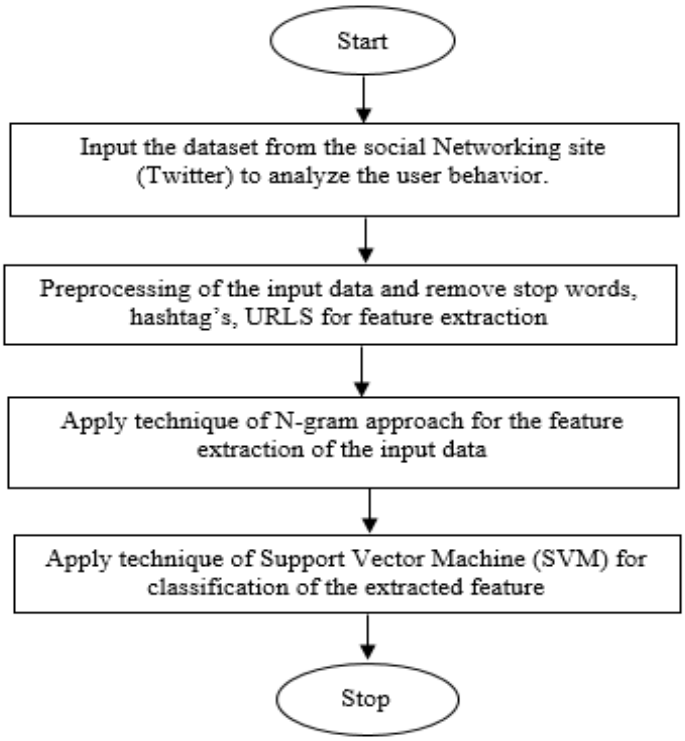

Fig.5. Proposed Flowchart

The flowchart depicts various steps that are involved in sentiment analysis technique.

CONCLUSION AND FUTURE WORK

The stupendous fact about social media is that it allows its users to freely share their experiences, reviews, thoughts, short messages and so on. In today's scenario, people life is incomplete without internet. Internet allows its users to express their opinion online. Written sentiment or opinion indicates the perspective of users on a specific entity. Analyzing these perspective is a challenging task when Homographs used. This paper discuss sentiment analysis its approaches and its techniques. N-gram technique is used for feature extraction and Support Vector Machine (SVM) technique used for Classification.

In this paper, a brief introduction on sentiment analysis is given. Machine learning approaches like supervised and unsupervised learning and Lexicon based approach like Dictionary and Corpus method has been discussed. Techniques like N-gram for feature extraction and SVM for classification on sentiments is presented. Further a workflow of sentiment analysis methods on input data set has been presented.

The research is based on the Sentiment Analysis of posts on Twitter using homographs. Presently it will work for English Language but in future can work in multilingual

Note:-This is a base line work for an on-going M.Tech Research work on Sentimental analysis on social media with reference to Twitter. The implementation of this work has already been started.

REFERENCES

[1] Fadhli Mubarak bin Naina Hanif, G. A. Putri Saptawati," CORRELATION ANALYSIS OF USER INFLUENCE AND SENTIMENT ON TWITTER DATA", 2014, IEEE, 978-1-4799-7996-7

[2] Zhou Jin, Yujiu Yang, Xianyu Bao, Biqing Huang," Combining User-based and Global Lexicon Features for Sentiment Analysis in Twitter", 2016, IEEE, 978-1-5090-0620-5

[3] M.Trupthi, Suresh Pabboju,"Sentiment Analysis on Twitter using Streaming API", 2017, IEEE, 978-1-5090-1560-3.

[4] F. Aisopos, G. Papadakis, and T. Varvarigou, "Sentiment analysis of social media content using N-Gram graphs," *Proc. 3rd ACM SIGMM Int. Work. Soc. media - WSM '11*, no. November, p. 9, 2011.

[5] Nehal Mamgain, Ekta Mehta, Ankush Mittal, Gaurav Bhatt," Sentiment Analysis of Top Colleges in India Using Twitter Data", 2016, IEEE, 978-1-5090-0082-1

[6] L. Auria and R. A. Moro, "1," no. August, 2008.

[7] A. kathuria and S. Upadhyay,"A Novel Review of various Sentimental Analysis Techniques," vol. 6, no. 4, pp.17-22, 2017

[8] Deepali Arora, Kin Fun Li and Stephen W. Neville," Consumers' sentiment analysis of popular phone brands and operating system preference using Twitter data: A feasibility study", 2015, IEEE, 1550-445X

[9] 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.

[10] Geetika Gautam, Divakar yadav," Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis", 2014, IEEE, 978-1-4799-5173-4

[11] K. Chang, "Lecture 2: N-gram," pp.1-45.

[12] Ryan M. Eshleman and Hui Yang," A Spatio-temporal Sentiment Analysis of Twitter Data and 311 Civil Complaints", 2014, IEEE, 978-1-4799-6719-3

[13] http://www.statsoft.com/Textbook/Support-Vector-Machines

[14] https://www.udacity.com/course/ud120

[15]https://stats.stackexchange.com/questions/312897/c-classification-svm-vs-nu-classification-svm-in-e1071-r

[16] http://www.svms.org/parameters/

[17] “Chapter 2 A Brief Introduction to Support Vector Machine (SVM),” 2011

[18] D.Hillard and S. Petersen, “N-gram Language Modeling Tutorial,” no. Lm, pp. 1–19, 2001.

[19]“N-Gram Model Formulas Laplace (Add-One) Smoothing Formal Definition of an HMM Computing the Forward Probabilities,” vol. 1.

[20] B. S. Dattu and P. D. V Gore, “A Survey on Sentiment Analysis on Twitter Data Using Different Techniques,” vol. 6, no. 6, pp. 5358–5362, 2015.

[21] Akankasha and Bhavna Arora" A Review of Sentimental Analysis on Social Media Application", abstract in proceeding of ICETEAS-2018,International Conference on Emerging Trends in Expert Application & Security,17-18,Feb,2018