International Journal of Advanced Research in Science Management and Technology

ISSN(Online)-2454-4159 Volume 6, Issue 1, January 2020

"Social media users" Sentiment Analysis by Naïve Bayes text mining algorithm

Mr.Ashish S. Awate[1],Mr. Bhushan N. Nandwalkar[2]

Assistant Professor. Department of Computer Engineering, SVKM's IOT, Dhule, Maharashtra, India[1,2]

Abstract— The errand of finding, looking, to remove and to arrange the conclusion on is named as Sentiment Analysis (SA). SA goes under the following of open sentiments for specific approaches, lawa, or promoting techniques by processing the natural language (NLP) It includes a way that improvement for the assortment and assessment of remarks and suppositions about enactment, laws, strategies, and so on., which are posted on the internet based life. This paper tends to the issue of sentiment analysis in twitter; that is grouping tweets as per the sentiment communicated in them: positive and negative. Twitter is an online miniaturized scale blogging and long range interpersonal communication stage which permits clients to compose short announcements of most extreme length 140 characters. It is a quickly extending administration with more than 200 million enlisted clients - out of which 100 million are dynamic clients and half of them sign on twitter regularly - producing almost 250 million tweets for each day. Because of this enormous measure of utilization we would like to accomplish an impression of open sentiment by breaking down the sentiments communicated in the tweets.

Breaking down the open sentiment is significant for some applications, for example, firms attempting to discover the reaction of their items in the market, anticipating political decisions and foreseeing financial wonders like stock trade. In this paper we build up a useful classifier for precise and programmed sentiment classification of an obscure tweet stream by utilizing Naïve Bayes and most extreme entropy calculation. Here we utilized a dataset shaped of gathered messages from Twitter in test arrangement. The objective is to mechanize the way toward mining mentalities, suppositions and concealed feelings from content.

Keywords— Sentiment Analysis, Naïve Bayes, Entropy, Classification, Natural Language Processing, Microblogging.

I. INTRODUCTION

Microblogging today has become a well known specialized instrument among Internet clients. A large number of messages are showing up day by day in mainstream sites that offer types of assistance for microblogging, for example, Twitter, Tumblr, Facebook. Writers of those messages expound on their life, share feelings on assortment of points and talk about current issues. As a result of a free arrangement of messages and a simple availability of microblogging stages, Internet clients will in general move from customary specialized apparatuses, (for example, conventional online journals or mailing records) to microblogging administrations.

As an ever increasing number of clients post about items and administrations they use, or express their political and strict perspectives, microblogging sites become significant wellsprings of individuals' assessments and sentiments. Such information can be proficiently utilized for showcasing or social investigations. We utilize a dataset framed of gathered messages from Twitter.

1.1. Inspiration

We have decided to work with twitter since we feel it is a superior guess of open sentiment instead of ordinary web articles and web sites. The explanation is that the measure of pertinent information is a lot bigger for twitter, when contrasted with conventional blogging destinations. In addition the reaction on twitter is increasingly expeditious and furthermore progressively broad (since the quantity of clients who tweet is significantly more than the individuals who compose web writes every day).

Sentiment analysis of open is profoundly basic in full scale financial marvels like anticipating the securities exchange pace of a specific firm. This should be possible by breaking down generally speaking open sentiment towards that firm as for time and utilizing financial aspects instruments for finding the relationship between's open sentiment and the association's securities exchange esteem. Firms can likewise appraise how well their item is reacting in the market, which territories of the market is it having a good reaction and in which a negative reaction [1].

On the off chance that organizations can get this data they can break down the explanations for separated reaction, thus they can showcase their item in a more improved way by searching for fitting arrangements like making appropriate market fragments. Anticipating the consequences of well known political races and surveys is likewise a developing application to sentiment analysis.

1.2 Domain Introduction

This breaking down sentiments of tweets goes under the space of "Example Classification" and "Information Mining". Both of these terms are firmly related and interlaced, and they can be officially characterized as the way toward finding "valuable" designs in enormous arrangement of information, either consequently (unaided) or semi-naturally (regulated). The models would intensely depend on methods of "Natural Language Processing" in removing critical examples and highlights from the huge informational collection of tweets and on "AI" strategies for precisely grouping individual unlabelled information tests (tweets) as indicated by whichever example model best depicts them. The highlights that can be utilized for displaying examples and classification can be isolated into two fundamental gatherings: formal language based and casual blogging based.

Language based highlights are those that manage formal etymology and incorporate earlier sentiment extremity of individual words and expressions, and grammatical features labeling of the sentence. Earlier sentiment extremity implies that a few words and expressions have a natural intrinsic propensity for communicating specific and explicit sentiments all in all. For instance "magnificent" has a solid positive undertone while "insidious" has a solid negative implication. So at whatever point a word with positive meaning is utilized in a sentence, odds are that the whole sentence would communicate a positive sentiment.

Grammatical features labeling, then again, are a linguistic way to deal with the issue. It intends to consequently distinguish which grammatical feature every individual expression of a sentence has a place with: thing, pronoun, modifier, descriptive word, action word, interposition, and so forth. Examples can be removed from dissecting the recurrence dissemination of these grammatical forms (ether exclusively or all things considered with some other grammatical feature) in a specific class of marked tweets. Twitter based highlights are increasingly casual and relate with how individuals communicate on online social stages and pack their sentiments in the restricted space of 140 characters offered by twitter. They

incorporate twitter hashtags, retweets, word, question marks, nearness of url in tweets, shout marks, web emojis and web shorthand/slangs [3].

II. LITERATURE SURVEY

With the number of inhabitants in sites and interpersonal organizations, conclusion mining and sentiment analysis turned into a field of enthusiasm for some explores. An expansive review of the current work was introduced in (Kumar Raviab and Vadlamani Rav, 2015). A wide review of the current work was introduced in this overview covering distributed writing during 2002–2015, is composed based on sub-undertakings to be performed, AI and natural language processing methods utilized and uses of sentiment analysis. Additionally (R.Piryania,D.Madhavib,V.K.Singh) in their paper introduced a point by point explanatory mapping of OMSA examine work and diagrams the advancement of order on different valuable parameters.Pang and Lee, 2008 in their study, the creators portray existing systems and approaches for an assessment situated data recovery. In any case, very few looks into in feeling mining thought about online journals and even significantly less tended to microblogging. In (Yang et al., 2007), the creators use web-sites to build a corpora for sentiment analysis and use feeling symbols relegated to blog entries as markers of clients' state of mind. The creators applied SVM and CRF students to group sentiments at the sentence level and afterward examined a few systems to decide the general sentiment of the archive. As the outcome, the triumphant technique is characterized by considering the sentiment of the last sentence of the report as the sentiment at the record level.

(Krishna B Vamshi ; Ajeet Kumar Pandey ; Kumar A. P. Siva 2018) talks about another subject model based methodology for feeling mining and sentiment analysis of content surveys posted in web discussions or online networking website which are generally in unstructured in nature.

J. Peruse in (Read, 2005) utilized emojis, for example, ":-)" and ":- (" to shape a preparation set for the sentiment classification. For this reason, the creator gathered writings containing emojis from Usenet newsgroups. The dataset was partitioned into "positive" (writings with glad emojis) and "negative" (writings with pitiful or furious emojis) tests. Emoticonstrained classifiers: SVM and Na¨ıveBayes, had the option to get up to 70% of an exactness on the test set.

Sentiment analysis has been dealt with as a Natural Language Processing task at numerous degrees of granularity. Beginning from being a report level classification task (Turney, 2002; Pang and Lee, 2004), it has been taken care of at the sentence level (Hu and Liu, 2004; Kim and Hovy, 2004) and all the more as of late at the expression level (Wilson et al., 2005; Agarwal et al., 2009). Microblog information like Twitter, on which clients present ongoing responses on and conclusions about "everything", presents fresher and various difficulties. A portion of the previous outcomes on sentiment analysis of Twitter information are by Go et al. (2009), (Bermingham and Smeaton, 2010) and Pak and Paroubek (2010). Go et al. (2009) utilize inaccessible figuring out how to secure sentiment information. They use tweets finishing off with positive emojis and negative emojis. They assemble models utilizing Naive Bayes, MaxEnt and Support Vector Machines (SVM), and they report SVM beats different classifiers [1].

Regarding highlight space, they attempt a Unigram, Bigram model related to grammatical features (POS) highlights. They note that the unigram model beats every other model. In particular, bigrams and POS highlights don't help. Pak and Paroubek (2010) gather information following a comparable far off learning

worldview. They play out an alternate classification task however: abstract versus objective. For emotional information they gather the tweets finishing with emojis in a similar way as Go et al. (2009). For target information they slither twitter records of well known papers like.

III. RESEARCH IDENTIFIACTION

Better than average measure of related earlier work has been done on sentiment analysis of client audits [x], records, web sites/articles and general expression level sentiment analysis. These vary from twitter primarily due to the furthest reaches of 140 characters for each tweet which powers the client to communicate assessment compacted in short content. The best outcomes came to in sentiment classification utilize administered learning strategies, for example, Naive Bayes and MaxEnt[11], yet the manual marking required for the managed approach is extravagant. Some work has been done on solo and semi-directed methodologies, and there is a great deal of room of progress. Different scientists were trying new highlights and classification methods frequently simply contrast their outcomes with pattern execution. There is a need of appropriate and formal correlations between these outcomes showed up through various highlights and classification procedures so as to choose the best highlights and most effective classification methods for specific applications [8].

Applying sentiment analysis on Twitter is the forthcoming need with analysts perceiving the logical preliminaries and its potential applications. The moves remarkable to this issue zone are to a great extent ascribed to the overwhelmingly casual tone of the miniaturized scale blogging. They use microblogging and all the more especially Twitter as a corpus for sentiment analysis. They refered to:

1. Microblogging stages are utilized by various individuals to communicate their supposition about various themes, in this way it is an important wellspring of individuals' assessments.
2. Twitter contains a colossal number of content posts and it develops each day. The gathered corpus can be subjectively enormous.
3. Twitter's crowd shifts from customary clients to VIPs, organization delegates, government officials, and even nation presidents. In this manner, it is conceivable to gather content posts of clients from various social and interests gatherings.
4. Twitter's crowd is spoken to by clients from numerous nations.

So we executed Naive Bayes ngram models and a Maximum Entropy model to characterize tweets. The analysts found that the Naive Bayes classifiers worked obviously superior to the Maximum Entropy model could. The preparation information comprised of tweets with emojis. The emojis filled in as boisterous names. So there was a need to fabricate models utilizing Naive Bayes, MaxEnt Their element space comprised of unigrams, bigrams and POS. The detailed that unigram were increasingly successful as highlights yet we utilized ngram as our standard.

The comparative works has been done yet arrange the tweets as target, positive and negative. So as to gather a corpus of target posts, they recovered instant messages from Twitter records of well known papers and magazine, for example, "New York Times", "Washington Posts" and so forth. Their classifier depends on the multinomial Naïve Bayes classifier that utilizes N-gram and POS-labels as highlights, excessively characterized tweets as target or abstract and afterward the emotional tweets were delegated positive or negative. The issues which are not considered about tweets like retweet, hashtags, connection, accentuation and outcry checks related to highlights like earlier extremity of words and POS of words [8].

International Journal of Advanced Research in Science Management and Technology
ISSN(Online)-2454-4159 Volume 6, Issue 1, January 2020

IV. RESEARCH METHODOLOGY

In the past decade, new forms of communication, such as microblogging and text messaging have emerged and become ubiquitous. While there is no limit to the range of information conveyed by tweets and texts, often these short messages are used to share opinions and sentiments that people have about what is going on in the world around them. Classify whether the message is of positive, negative sentiment. For messages conveying both a positive and negative sentiment, whichever is the stronger sentiment should be chosen.

Working with these informal text genres presents challenges for natural language processing beyond those typically encountered when working with more traditional text genres, such as newswire data. Tweets and texts are short: a sentence or a headline rather than a document. The language used is very informal, with creative spelling and punctuation, misspellings, slang, new words, URLs, and genre-specific terminology and abbreviations, such as, RT for "re-tweet" and # hashtags, which are a type of tagging for Twitter messages. How to handle such challenges so as to automatically mine and understand the opinions and sentiments that people are communicating has only very recently been the subject of research [3,6].

Proposed System architecture

Our system is organized in five main components: preprocessing of tweets, feature extraction, training set which is a set of predefined positive or negative tweets used for building the sentence database against it the classification of a query tweet is done, classifier using naive Bayes or , MaxEnt and the output(positive or negative).

These components are connected in pipeline architecture, shown in figure (1). The classifier determines the polarity class of the tweet message as a final output.

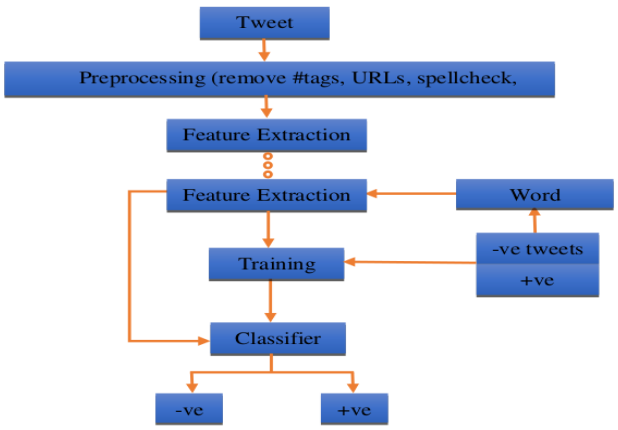


Fig.1 System architecture

Here it's presented a method which collects a corpus with positive and negative sentiments. The method allows collecting negative and positive sentiments such that no human effort is needed for classifying the documents. The size of the collected corpora can be arbitrarily large but we limited the size. Secondly we perform statistical linguistic analysis of the collected corpus. Thirdly we use the collected corpora to build a sentiment classification system for microblogging in two categories. Finally we conduct

experimental evaluations on a set of real microblogging posts to prove that the techniques used are efficient and perform better than previously proposed methods.

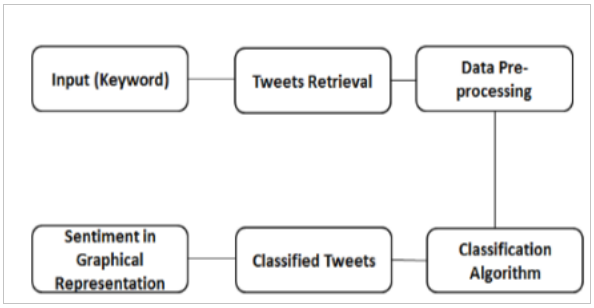


Fig.2 n-gram

Construction of n-grams

In the fields of computational linguistics and probability, an n-gram is a contiguous sequence of n items from a given sequence of text or speech. The items can be phonemes syllables, letters, words or base pairs according to the application. The n-grams typically are collected from a text or speech corpus. When the items are words, n-grams may also be called shingles.

The baseline of implemented model is n-gram. An n-gram of size 1 is referred to as a "unigram"; size 2 is a "bigram" (or, less commonly, a "digram"); size 3 is a "trigram". Larger sizes are sometimes referred to by the value of n, e.g., "four-gram", "five-gram", and so on.

Set of n-grams can be made out of consecutive words. Negation words such as “no”, “not” is attached to a word which follows or precedes it. For example: “I do not like soda” has two bigrams: “I do+not”, “do+not like”, “not+like soda”. So the accuracy of the classification improves by such procedure, because negation plays an important role in sentiment analysis [4,7].

Classifiers

Two different classifiers are used here. A Naive Bayes classifier and Maximum Entropy[1,2].

Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

Another classifier the Max Entropy classifier is a probabilistic classifier which belongs to the class of exponential models. Unlike the Naive Bayes classifier that we discussed in the previous article, the Max Entropy does not assume that the features are conditionally independent of each other.

The MaxEnt is based on the Principle of Maximum Entropy and from all the models that fit our training data, selects the one which has the largest entropy. The Max Entropy classifier can be used to solve a large variety of text classification problems such as language detection, topic classification, sentiment

analysis and more [6].

Corpus Collection

Using Twitter API we collected a corpus of text posts and formed a dataset of two classes: positive sentiments and negative sentiments. To collect negative and positive sentiments, we followed the procedure in which we fetch some limited tweets by using twitter API .We queried Twitter for two types of emoticons:

a. Happy emoticons: ":-)", ":)", "=)", ":D" etc.

b. Sad emoticons: ":-(", ":(", "=(", ";(" etc.

The two types of collected corpora will be used to train a classifier to recognize positive and negative sentiments. Each message cannot exceed 140 characters by the rules of the microblogging platform; it is usually composed of a single sentence. Therefore, we assume that an emoticon within a message represents an emotion for the whole message and all the words of the message are related to this emotion. We use English language. However, the method can be adapted easily to other languages too since Twitter API allows to specify the language of the retrieved posts.

Pre-processing of tweets data and feature extraction

After arriving at training set we need to extract useful features from it which can be used in the process of classification. Following text formatting techniques gets aid in feature extraction:

a. **Tokenization:** It is the process of breaking a stream of text up into words, symbols and other meaningful elements called "tokens". Tokens can be separated by whitespace characters and/or punctuation characters.

b. **Url's and user references** (identified by tokens "http" and "@") are removed if we are interested in only analyzing the text of the tweet.

c. **Punctuation marks and digits/numerals** may be removed if for example we wish to compare the tweet to a list of English words.

d. **Lowercase Conversion:** Tweet may be normalized by converting it to lowercase which makes it's comparison with an English dictionary easier.

e. **Stemming:** It is the text normalizing process of reducing a derived word to its root or stem. For example a stemmer would reduce the phrases "stemmer", "stemmed", "stemming" to the root word "stem". Advantage of stemming is that it makes comparison between words simpler, as we do not need to deal with complex grammatical transformations of the word. In our case we employed the algorithm of "porter stemming"[8] on both the tweets and the dictionary, whenever there was a need of comparison.

f. **Stop-words removal:** Stop words are class of some extremely common words which hold no additional information when used in a text and are thus claimed to be useless. Examples include "a", "an", "the", "he", "she", "by", "on", etc. It is sometimes convenient to remove these words because they hold no additional information since they are used almost equally in all classes of text, for example when computing prior-sentiment-polarity of words in a tweet according to their frequency of occurrence in different classes and using this Project Thesis Report 29 polarity to calculate the average sentiment of the tweet over the set of words used in that tweet.

g. **Parts-of-Speech Tagging:** POS-Tagging is the process of assigning a tag to each word in the sentence as to which grammatical part of speech that word belongs to, i.e. noun, verb, adjective, adverb,

coordinating conjunction etc.

Twitter data are prepared with the aid of two dictionaries, i.e., emoticon dictionary and acronym dictionary. The emoticon dictionary contains about many labelled emoticons ":)" labelled as positive and ":)" labelled as negative. The acronym dictionary has many acronyms, e.g., WOW (Wonder of Wonders).

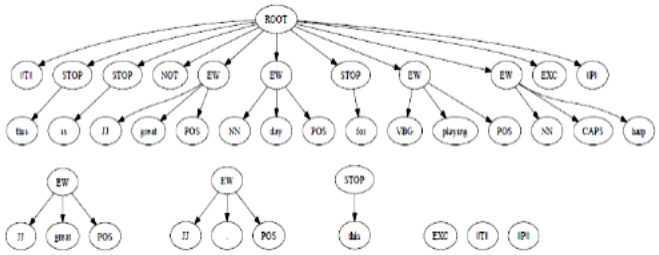


Fig.3. Tree kernel for a tweet: "@Fernando this isn't a great day for playing the HARP! :)"

The given tweet will be handle as follows:

1. Look up for emoticons and their sentiment polarity (positive, negative) in emoticons dictionary.
2. Replace all URLs with a tag ||U||.
3. Replace targets (e.g. "@Fernando") with tag ||T||.
4. Replace all negations by tag "Not".

After performing the former steps, we remove hashtags, URLs, and make spell check. The next step is to make emoticons tagging and POS(part-of-speech) tagging, POS tagging is the most difficult part, as one have to assign it to each word in a sentence. For example, a sentence like "Heat water in a large vessel" will be [heat(verb) water(noun) in(prep.) a(det.) large(adj.) vessel(noun)] words associated with their tags. In the final step, POS used to build a sentence database, namely, verbs, adjectives, emoticons, etc. In information retrieval (IR), POS used to compute a term weights, which are mathematical computations of how informative words are, and constitute an integral part of the statistical modelling of documents by IR systems [3,6].

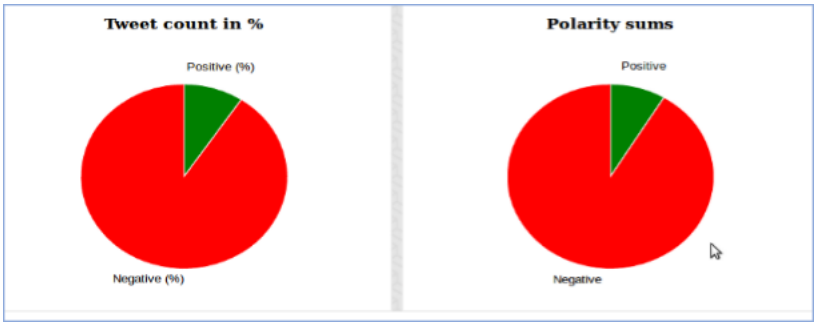
Training Data

In order to train a supervised learning algorithm a training dataset must be collected; this dataset consists of training examples and the corresponding expected output for each example. The expected output is general known as the target. A supervised learning algorithm uses this dataset so that it can learn to map the input examples to their expected target. If the training process is implemented correctly the machine learning algorithm should be able to generalise the training data so that it can correctly map new data that it has never seen before. Training data must contain a class label, this can be achieved through manually assigning each tweet with a class but this is a tedious process and as twitter enforces strict rules on the distribution of its data it has proved difficult to source reliable hand annotated twitter datasets. In order to use this method an assumption must be made, this assumption is that the emoticon in the tweet represents the overall sentiment contained in that tweet. This assumption is quite reasonable as the maximum length of a tweet is 140 characters so in the majority of cases the emoticon will correctly represent the overall sentiment of that tweet. In this we used the smiley face and the sad face

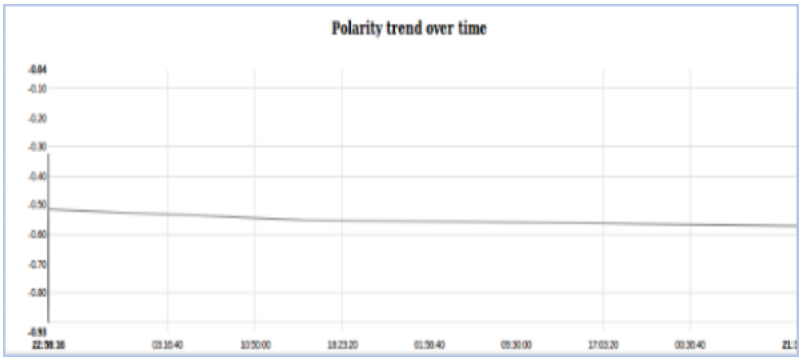
were chosen as the noisy labels for the trainingdata, this choice was made as they were the two labels with the highest frequency that represented either the positive or negative class [7,10]

V. DATA ANALYSIS

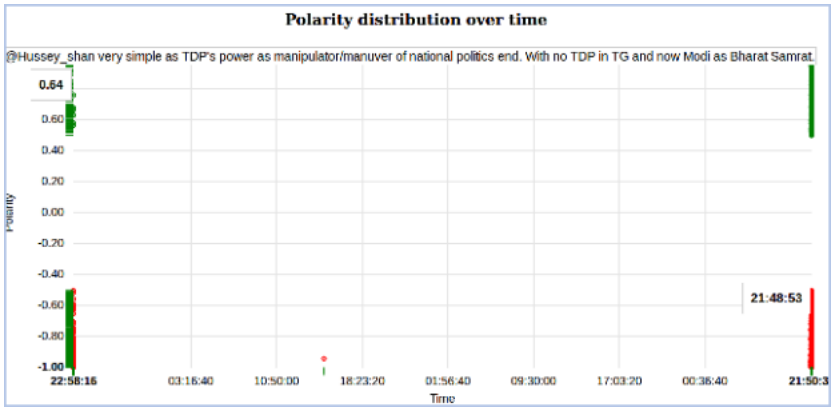
➢ After inputting a keyword the corresponding output is shown.



Graph 1- The calculated polarity of collected tweets is shown. The polarity is shown in both ways like tweet count in % and polarity sums.



Graph 2- As per the variation in time the polarity gets change. The polarity trend over time is also crucial parameter to be considered.



Graph 3- Polarity distribution over time.

VI. CONCLUSION

Microblogging nowadays became one of the major types of the communication. The large amount of information contained in microblogging web-sites makes them an attractive source of data for opinion mining and sentiment analysis. In this wementioned a method for an automatic collection of a limited

corpus that can be used to train a sentiment classifier. We constructed Tree kernel for POS-tagging and observed the difference in distributions among positive, negative sets. From this we conclude that syntactic structures to describe emotions or state facts. Some POS-tags may be strong indicators of emotional text and by using different NLTK classifier it is easier to classify the tweets and more we improve the training data set more we can get accurate results.

We used the collected corpus to train a sentiment classifier. Our classifier is able to determine positive and negative sentiments of documents. The classifier is based on the multinomial Naıve Bayes classifier that uses N-gram and POS-tags as features. Variation in Polarity as with variation in time is shown in polarity distribution over time it also includes the corresponding tweet as the cursor is pointed at particular green (positive tweet) and red (negative tweet) dot. We can improve the implemented model by adding extra information like closeness of the word with a negation word. The closer the negation word is to the ngram word whose prior polarity is to be calculated, the more it should affect the polarity. For example the position of negation word, it may simply reverse the polarity of that word and farther the negation is from the word the more minimized its effect should be. The issues which are not considered about tweets like retweet, hashtags, link, punctuation and exclamation marks in conjunction with features like prior polarity of words and POS of words can be considered while analysing sentiments. We are classifying tweets, either as positive or negative sentiment.

REFERENCES

[1] Krishna B Vamshi ; Ajeet Kumar Pandey ; Kumar A. P. Siva, 'Topic Model Based Opinion Mining and Sentiment Analysis', 2018 International Conference on Computer Communication and Informatics (ICCCI)

[2] Piryania D, Madhavib, V.K.Singh 'Analytical mapping of opinion mining and sentiment analysis research during 2000–2015' Information Processing & Management Volume 53, Issue 1, January 2017, Pages 122-150

[3] Kumar Raviab, Vadlamani Ravia 'A survey on opinion mining and sentiment analysis: Tasks, approaches and applications', Scince Direct Journal Knowledge-Based Systems Volume 89, November 2015, Pages 14-46

[4] Alexander Pak and Patrick Paroubek. 'Twitter as a corpus for sentiment analysis and opinion mining' Conference: Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta.

[5] Wilson, J. Wiebe, and P. Hoffman. 'Recognizing contextual polarity in phrase level sentiment analysis.' HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language ProcessingOctober 2005 Pages 347–354

[6] ApoorvAgarwal, FadiBiadsy, and Kathleen Mckeow,. 'Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams'. EACL '09: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics

[7] Alec Go, RichaBhayani and Lei Huang. Twitter Sentiment Classification using Distant Supervision. Project Technical Report, Stanford University, 2009.

[8] EfthymiosKouloumpis, Theresa Wilson and Johanna Moore. 'Twitter Sentiment Analysis: The Good the Bad and the OMG!' In Proceedings of AAAI Conference on Weblogs and Social Media (ICWSM), 2011.

[9] Barbosa, L., and Feng, J. 2010. 'Robust sentiment detection on twitter from biased and noisy data. Coling 2010: Poster Volume, pages 36–44, Beijing, August 2010

[10] Bifet, A., and Frank, E. 2010. 'Sentiment knowledge discovery in twitter streaming data' DS'10: Proceedings of the 13th international conference on Discovery scienceOctober 2010 Pages 1–15

[11] Davidov, D.; Tsur, O.; and Rappoport, A.. 'Enhanced sentiment learning using twitter hashtags and smileys' COLING '10: Proceedings of the 23rd International Conference on Computational Linguistics: PostersAugust 2010 Pages

241–249

[12] Esuli, A., and Sebastiani, F. 2006. 'SentiWordNet: A publicly available lexical resource for opinion mining.' Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06) May 2006 Genoa, Italy