International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-1, November 2019

Predicting Election Results using NLTK

Kambhampati Kalyana Kameswari, J Raghaveni, R. Shiva Shankar, Ch. Someswara Rao

Abstract: ***In today's world, people are usually using social media networks for trying to communicate with other users and for sharing information across the world. The online social networking sites have become considerable tools and are providing a common medium for a number of users to communicate with each other. Twitter is the most prominent microblogging website and one among the social networking sites that grow on a daily basis. Social media incorporates an extensive amount of data in the form of tweets, forums, status updates, comments, etc. in an attempt to automatically process and analyze these data, applications can rely on analysis approaches such as sentiment analysis. Twitter sentiment analysis is an application of sentiment analysis on data from Twitter (tweets), to obtain user's opinions and sentiments. Natural Language Toolkit (NLTK) is a library based on machine learning methods in python & sentiment analysis tool. Which provides the base for text processing and classification? The research work proposed a machine learning-based classifier to extract the tweets on elections and analyze the opinion of the tweeples (people who use twitter). The tweets can be categorized as positive, negative and neutral towards a particular politician. We classify these processed tweets using a supervised machine learning classification approach. The classifier used to classify the tweets as positive, negative or neutral is Naive Bayes Classifier. The classifier is trained with tweets bearing a distinctive polarity. The percentage of positive and negative tweets is then measured and graphically represented.***

Keywords: Machine Learning, Natural Language Processing, Twitter, Political opinion, Supervised Learning, Election Result Prediction.

I. INTRODUCTION

Elections play a major role in democracy. Elections enable citizens to elect their leaders. It provides everyone a chance for equal representation and voice in our government. Democracy is the government for the people, and by the people, which means government leaders are determined by participation in elections. It is the primary democratic tool in which the people interact with the representatives. Because of its significant position in politics, the prediction of an election result was always a major concern. One important consideration in the election is that the polls/surveys.

Since 1824, polls were being used to take a snapshot of public opinion, Public opinion polls have been conducted on every subject under the sun, from President's approval to celebrities ' events and sport forecasts, but are particularly important for conducting and analyzing elections. Election results and opinion polls are so interconnected that it is hard to envision one without the other. Poll ratings provide political support for media coverage and election predictions, they frame nominee and voter behavior, and they are the justification of interpreting the meaning of election outcomes [5].

But even in developed countries, polling often did not predict the electoral results correctly. A number of lost votes were included, such as the 1992 British General Elections, the 1998 Quebec Elections, the 2002 and 2007 French Presidential elections, the 2004 European election in Portugal, the 2006 Italian Common Elections, and the 2008 Primary Elections in the States [2].

For years, business scientists have applied standardized methodologies like surveys to assess individual groups' views and motives, which have several drawbacks, including the human endeavor engaged and may be time-consuming and expensive.

Recently, it is noted that standard polls can sometimes fail to make an accurate prediction. As an alternative way of estimating election results, the scientific community has hopefully switched its attention to analyzing internet information such as blog posts or user activities on the social network. Moreover, traditional surveys are too expensive and internet data is simple to acquire and easily approachable hence to determine the accuracy and high-cost problem, we consider the possibility of using data from online social networks as the data source to predict the outcome of an election. [2].

Social media has always been a leading component of most people's lives and it's been changing an aspect of their lives in many ways since the last few years. It has become one of the key media of internet communication. People publish their opinions on a variety of topics and discuss the latest trends, post their daily life activities. Social media like Twitter, My Space, Facebook, Instagram. Platforms like LinkedIn, Facebook, Twitter, Instagram, My Space, Tumbler and Google+ are being profoundly used to share opinions, reviews, suggestions, and ratings [12].

Twitter is a social media network that allows researchers to use their data. Twitter is a microblogging web service that was introduced in 2006. At present it reaches 200 million users per month and 500 million posts per day. Users of Twitter are allowed to post up to 140 message characters (tweet).

Revised Manuscript Received on November 30, 2019.
***** Correspondence Author
Kambhampati Kalyana Kameswari*, M. tech Student, Department of CSE, SRKR Engineering College affiliated to JNTU Kakinada, Bhimavaram, AP, India. Email: kalyani95.kambhampati@gmail.com
J Raghaveni, Assistant Professor of Computer Science and Engineering, SRKR Engineering College affiliated to JNTU Kakinada, Bhimavaram, AP, India. Email: hariveni9@gmail.com
R.Shiva Shankar, Assistant Professor of Computer Science and Engineering, SRKR Engineering College affiliated to JNTU Kakinada, Bhimavaram, AP,India. Email: shiva.srkr@gmail.com
Ch. Someswara Rao, Assistant Professor of Computer Science and Engineering, SRKR Engineering College affiliated to JNTU Kakinada, Bhimavaram, AP, India. Email: chinta.someswararao@gmail.com

Twitter is an online platform that allows users to write short status updates on microblogging and social networks. It is an extensive and growing network of over 200 million unique users, 100 million of which are active users, half of whom log on twitter-generating approximately 250 million tweets a day. We expect to represent public sentiments through examines of the emotions expressed in our tweets. Public opinion analysis is important for many purposes, like companies trying to determine the reaction of their products on the market, predicting political elections, predicting political / economic trends like the stock exchange etc. [15].

Most of the people rely heavily on online information published by users for decision-making. For example, If anyone wants to buy or use an item, they first check at their reviews online before making a decision, then address it on the social media. The amount of user content is too large to be analyzed by a normal user. This desperately wants to be automated, multiple techniques of sentiment analysis are extensively used to analyze the data.

Sentiment analysis (SA) shows users whether the product information is adequate or not before purchasing it. Marketers and companies use this analysis data in order to understand their product or service so that it can be offered according to the needs of the user. The recovery methods for textual information concentrate primarily on the processing, search or analysis of the actual data observed. The facts have an objective component, but other texts express subjective characteristics. This information is essentially a center of sentiment analysis (SA) views, feelings, assessments, and perceptions. It provides many difficult possibilities for developing innovative applications, primarily because of the enormous increase in data accessible on internet outlets such as blogs and social networks.[4].The sentiment analysis can be described as a tool that automates the exploitation by means of Natural Language Processing (NLP) of behaviors, views, and emotions in text, speech, posting on social media, and database data. Analysis of sentiments includes the classification of views in writing into "positive"/" negative" / neutral "classifications. It is also related to as assessment of subjectivity, opinion mining and evaluation mining.

Application of opinion mining on these social media data was seen by many as a powerful tool to track user preferences and requirements. We are attempting to provide the general public with one such platform, one that is specific to political discussions and political debates. Political parties can also exploit the data collected on this platform to get insights into the feelings of the customers and therefore schedule their political campaigns. Political parties can also gain a brief insight into their possibility of winning the election. Furthermore, huge funds are poured out by political parties throughout elections for social media campaigns. The system we have created focusses on assigning every post published by its users to the continuing discussion or debate on the political parties using common text classification algorithms such as Naive Bayes which convey their views, emotions, feelings, etc., in a supervised Learning Algorithm.

This paper focuses in general on the US presidential elections set for 8 November 2016. The objective was to collect tweets referring to the elections and in particular to the two main candidates: Clinton and Trump. After our data have been acquired the method proposed is a selection and implementation of the classification algorithm. The key term is' sentiment analysis' (or opinion mining), in order to achieve text classification In this context, we present our forecast of the election results based on the approach suggested at the outset of this thesis.

II. LITERATURE SURVEY

Namrata Godbole et al.[1] Proposed the large-scale sentiment assessments for media and blogs, revealed a scheme consisting of a feeling identification stage, a sentiment accumulation, and a feeling screening stage that ranks every object compared to other individuals in that same category. The frequency of adjectives is monitored using WordNet with positive and negative polarity. Sentiment hop is necessary to determine the strength and eradicate ambiguous terms of the candidate outcomes. Adjectives divided by "And" provide the same polarity, but the adjectives divided by "But." have opposite polarity. Machine training techniques are more precise and execute faster than simple counting methods. Finally, they assessed the importance of scoring methods for a huge array of media and blogs. Godbole et.al concluded that sentiment may differ according to population group, news source or location. Mapping should be done to gain the best possible opinion using corpus-based techniques. In large scale data, it is complicated to track the opinions as well as owing to the different combinations of the phrases

Hailong Zhang et.al.[2] provides an analysis and comparison research with cross-domain, multi-lingual methods and certain evaluation methods of current opinion-mining techniques including computer training and lexicon-based approaches. Research findings demonstrate that machine learning techniques namely SVM and Naive Bayes are extremely accurate and the basic learning techniques are considered to be highly sustainable, while Lexicon-based methods requiring very little practice in documenting the human labelling process. In an attempt to resolve this problem, deep learning techniques are also being studied. Zhang et.al concluded that in prospective work, the research will concentrate on the combination of computer learning to view the lexicon method in order to strengthen the accuracy of the ranking of sentiments and the adaptability to distinct fields and literature.

Boia et al.[3] Experimented with tweets emoticon labelling. They suggest that for tweets with neutral sentiments that either got a positive or negative tag, most of the incorrect labelling was made. This demonstrates that an emoticon-based tag makes a very well-distinguished distinction between positive and negative emotions. Brief casual papers, such as tweets, blogs, and remarks, examine emotions as a principal element of data. Wishful emoticons and emoticons were found to be more commonly used

K. Manuel et al.[4]. developed a program for supervising Internet slang for sentiment analysis in blogs and review websites. A technique for calculating sentiment ratings for freshly discovered slang words were also provided in this work Identifying slang words and their meaning were first defined. Slang words are also used to evaluate feelings.

The writing is classified into subjective and objective. Using subjective phrases, the slang of sentiment is recognized. The score's polarity is determined using weighted reverse text frequency suggested two methods that depend on emoticons to identify the polarity of tweets and slang words to add a sentiment rank to internet documents, respectively. These two plays have shown how to use non-textual elements to identify a text's polarity. They noted that unigrams are the finest evaluation for this framework

Akcora et al.[5], tried to identify the emotional pattern and the word pattern that claims to change the public opinion, using Twitter data. To define the breakpoint, researchers use Jaccard's similarity of two successive intervals of words suggested a system that is able to Identify breakpoints in the public opinion that are used to determine how the public opinion varies over time. They track changes in the frequency of word use on social media, based on the idea that during an ongoing event, there are changes to the topics discussed, and therefore also to which words are used. Again, the work represented looks at opinions about individual entities and is, therefore, able to detect more fine-grained events

W Gao et.al [6] Contend that almost all previous research on Tweet Sentiment Classification is using a suboptimal methodology. The possible explanation is that most of these studies ' primary goal is not to estimate the class label (e.g. Positive, Negative, or Neutral) of every tweets, but to estimate the relative frequency (i.e. "prevalence") of the multiple classes in the dataset. The successive task is called quantification, and the latest research has convincingly demonstrated that it should be resolved as a problem on its own, using learning algorithms and evaluation tools other than those used for classification. Among this work, they clearly indicate, on a wide variety of TSC datasets, that using a quantification-specific algorithm generates considerably stronger category wavelength predictions than a frequently utilized advanced classification-oriented algorithm in TSC. They assert that the professionals involved should use quantification specific methodologies and designs for assessment instead of classification-specific ones.

Purtata Bhoir et al.[7] implemented a method for sentiment analysis of aspect level. The structure suggested applies to film review information. the information is pre-processed and POS marked originally. The reviews are classified with Naïve classifier and Senti word net as descriptive and subjective phrases. Subjective phrases are the phrases containing their views. The classification is initially trained with five thousand objective and five thousand subjective phrases. these pairs of features are then added to find a strong impression at the aspect level. As the final phase, Outlined polarity is generated for various aspects such as songs story, etc.

S Mandal et.al.[8] refers regarding to the internet where the growing popularity of web had shown the way to massive number of people to connect with one another beyond space and time this resulted to the enormous use as a forum for sharing views, exchanging thoughts, raising concerns, disrespect and so many other kinds of assessments of digital critics. thus the web turns into an enormous database of linguistic information that can be categorized to understand online user feelings and emotional state There are numerous algorithms for text classification proposed by researchers who take text s from online media and predict the understanding of a user after preprocessing, mining and classification of text. In this context S Mandal et.al proposed a Lexicon based text classification algorithm that is used to evaluate and anticipate a user sentiment polarity viz. Online assessments are positive, negative & neutral. In the context of three degrees of comparison factors, the proposed algorithm varies from the other Lexicon-based algorithms i.e positive, negative and neutral sentence grades for every word positive or negative. In addition, negation phrases demonstrate how system performance can be enhanced.

K Z Aung et.al.[9] Lexicon-based sentiment analysis for assessing teaching performance levels from textual feedback comments of students. A list of English words of sentiment is designed to classify the lexical origin of term polarity. Our Sentiment Word List contains the educational field perception terms to get better results. The value is given to each opinion in the database. The response value is between-3 and+ 3. This research suggests a Lexicon-based approach for the stage of teaching appraisal process. This approach correctly analyses participant reviews to be highly negative, moderately negative and weakly negative, or strongly positive, or moderately positive, or weakly positive or neutral category using two lexicons. or strongly positive, or moderately positive, or weakly positive or neutral classification using two lexicons. The degree of sentiment results for any instructor is given out from students' feedback responses.

Fumagalli et al.[10] listed several failure polls , such as the UK General Elections in 1992, the Quebec election in 1998 and the French Presidential elections in 2002 and 2007, the Portugal European elections in 2004, the 2006 Italian General Elections and the 2008 Primary State Elections. In developing countries such as Indonesia, this phenomenon is still happening, our records show that most of the polls in the 2012 Jakarta (Indonesia's capital city / province) governor election, the 2013 Bandung (West Java capital city) major election, and 2014 General elections, failed to predict the winner or have a large gap between the forecasts and the election results. Fumagalli et.al suggested using statistical matching and weighting methods to deal or to accommodate sampling of non-random objects.

R Rezapour et.al.[11] proposed and examined an enhanced model that integrates informative hashtags into a lexicon to increase the effectiveness of sentiment analysis. Alternatively, others examined the efficacy of tweets ' lexicon-original methods (LBA). In specific, Twitter information gathered about each of the Presidential applicants were analyzed with info on the hashtags and emoticons. Predicated on the notion that hashtags are informative terms and concatenated short phrases which contribute to conveying tweet feelings, therefore R Rezapour et.al have tested whether it improves sentiment predictive performance by incorporating prevalent hashtags from one dataset into a sentiment lexicon.

The results indicate that applicants can provide insightful information about the success of a candidate through the feeling of the tweets mentioned by these individuals-at least in social media. Therefore the work is limited.

Jyoti Ramteke et al.[12] first collected the data using the Twitter API and stored them in CSV file, then pre-processed it to remove specific characteristics and URLs and then hand-held data marking using the Hashtag Label, and then a VADER tool based on the Lexicon and the Regulative Sentiment Analysis Tool and introduced a scalable machine learning model to predict the election outcomes using two-stage frameworks to create training data from twitter data without negotiating on features and context. The sentiment analysis usually takes place on three stages: document-based, sentence-based and aspect-based. At the document-based, the whole document is taken into account, for example, say, a film is an entity and the whole document expresses a positive, negative or neutral polarity about a film review. All of the previously reviewed articles were document-based. Many data sets from product reports to hotel research are accessible for this task, but two of these well-researched data sets are discussed.

Mondher Bouazizi et.al.[13].presented the current research in order to quantify the various feelings within each tweet, which are very restricted in duration. Although tweets are permitted to only be 140 characters, they are most often emotional and have more than 1 sentiment. The text classification relates to the weight identification suggested for a sentiment analysis approach. Depending on the weight, the document is focused on positive, negative and mixed conditions and the accuracy of the information is 81%. In the next step, the sentiment quantification of tweets will be carried out by defining five positives and five negative subclasses of existing tweets A meaningful task and multi-class classification are performed For this reason the researchers suggested to carry out the ternary identification and then quantification in tweets with a range of pattern-based characteristics and unique Unigram-based characteristics along with other specific features. The researchers The primary goal is an efficient method for calculating the performance of the quantification and to assess and demonstrate the results of the methodology suggested..

Soler et al.[14], Provided a new tool for a Twitter assessment called Tara Tweet. By Incorporating this method 500.000 tweets had been assessed and the outcomes were shown at Tara Tweet's site. The findings of all three experiments are fairly accurate. Several parties have obtained various polls than was traditionally expected, which is natural, because votes can consider until the very last minute. With these experiments, a correlation between references and actual voting intentions can be identified. Parties invested in the promotion of social media will be more likely to see successful results in the polls, which suggest that Twitter analyzes are a secure way to carry out tests and produce results which are really similar to voter preferences in real time.

Java et al.[15] have shown that developers of social networking platforms predominantly post the results about their social interactions and perspectives, which makes social networks the most extensible and varied source of in-situ information about individuals's daily activities. They also analyzed an enormous social network on a different social media platform called microblogging. Such networks had a high degree of correlation and reciprocity, indicating that users had close mutual knowledge. While evaluating the purpose of an individual user to use such programs, we may clarify the aim of the Group by examining the aggregation behaviour of user groups. Understanding these intentions and learning how and why people are using these tools can be useful to improve and add new features that retain more users. We defined different types of user expectations and examined group mechanisms in this work They are currently working on predictive methods to identify consumer desires for associated social frameworks

III. METHODOLOGY

A. Data Collection (Tweets)

The data collection process is the first phase of the research, where twitter data is collected. The data retention process is challenging. In order to achieve the integrity and accuracy of the final outcomes, the collected data must be reliable and pre-processed properly. In some past document research, they developed a program to automatically gather a set of tweets relying on two kinds of tweets, "positive" and "negative." Twitter has two kinds of emoticons:

- Happy emoticons, such as ":)", ":P", ":)" etc.
- Sad emoticons, such as ":(", ":'(", "=(".

Many researchers are gathering and digitally annotating their own tweet dataset that is extremely lengthy and annoying. Likewise, to discover a means of obtaining a corpus of tweets, we must have a flexible data collection, which means we must have the same amount of positive and negative tweets as well as adequately wide. In fact, the more information we have, the more accurately we can train our classifier

After many kinds of research, I found a dataset of tweets in English coming from the source: Data world, It is composed of three columns that are ID, Text, and Source. We are only interested in the Text column (since the sentences were shorter than 160 characters and they were extracted from social media, we used them as tweets) containing the tweets in a raw format.

B. Data- preprocessing

Once the data was collected from twitter the next course of action is pre-processing that is implemented in python. For extracting features, data acquired by twitter is not suitable. Almost all tweets comprise of text including user names, empty space, special character, stop words, emoticons, acronyms, hashtags, metadata, URL"s, etc. So we use multiple NLTK.IN pre-processing functions to create this information suitable for extraction, we first extract our primary message from the tweet. then we eliminate all empty spaces, stop words(like is, a, the, he, them, etc.), hashtags, repeating words, URL"s, etc. we then replace all emoticons and abbreviations with their resultant meanings like,=D,=), lol, Rolf, etc.

Have been substituted with happy or giggle. Once we have accomplished this, we are willing for the necessary outcomes with our handled tweet. Tweet sampling and tweet processing.

Twitter information must be filtered because tweets involve several syntactic characteristics that might not be helpful to analyse.

C. Text Mining

Text extraction tends to refer to the analysis of data embedded in natural language text, (e.g. texts obtained from twitter). It can be determined as the procedure of synthesizing constructive knowledge from unorganized text sources. The implementation realm of text mining differs from clinical research applications to marketing applications and sentiment analysis. For the evaluation of customer relationship management, text mining is important in marketing. This way a company can improve its predictive analytics models for customer turnover (keep track of customer opinions). The primary objective of text classification is to process information into a structured format prepared for analysis, via the application of natural language processing and other theoretical frameworks. There are several aspects throughout the area of study of text mining, information extraction (IE) is appropriate for this endeavour. Ultimately, the preceding material intends to illustrate the challenges and terminology affiliated with information extraction and eventual monitoring.

D. Natural Language Processing

Twitter will be used as an instance to explain further notions. The information extracted from twitter presents a certain quantity of structuring, in the sense that perhaps the maximum limit of a tweet is 140 characters long. The duration restriction benefit is expressed in the evaluation difficulty for a single part of the text. However, the objective of the venture is to continuously analyze the information in which huge quantities of information are analyzed (e.g., 200 tweets a minute). In addition, no guarantee exists that each tweet follows a structured framework nor is it grammatically precise. However, sentences depicting the same or similar thoughts seem to have very distinct sentence construction and hire very specific vocabularies. Owing to the above structural constraints, a predetermined text structure is mandated at the moment of retrieval. The techniques portrayed below were used for the planning process.

E. Python

Python is a high standard, interpreted programming language, created by Guido van Rossum. For the readability, and compact line of codes, the language is very popular. It uses white space inundation to formalize blocks. Python provides a massive standard library that could be used for various applications, for instance, natural language processing, artificial intelligence, data analytics and so forth. It is prioritized for ambitious projects, because of its uniqueness, heterogeneous range of characteristics and its cohesive nature.

a. Natural Language Toolkit

Text classification is nothing but processing the extracted data before assessment. It involves the verification and termination of non-textual content and the content non-relevant to the field of data analysis. The best approach that could be used in data preprocessing is Natural Language Processing (NLP). In the current scenario, we import the database of NLTK (Natural Language Tool Kit). The Natural Language Tool Kit has become one of the finest-known and perhaps most-used NLP libraries in the Python ecosystem, beneficial for all various types of processes from tokenization, to stemming, to part of speech tagging, and beyond. text-processing libraries, NLTK offers easily used tools for the Ranking, Tokenization, Stemming, Tagging, Parsing and Semantic Analysis of NLP libraries, as well as a series of texting libraries for industry-starved applications.

All of the tweets are preprocessed by passing through the following steps in the same order.

- ***Tokenization:***

  The first objective, that should be performed before any processing may appear, is to split the textual data into individual components. This is a popular move in an implementation called Natural Language Processing (NLP). At a higher level, the text is divided into paragraphs and phrases. Due to twitter's 140-character duration restriction, it may indeed be the situation that more than one sentence is present in a tweet. In this aspect, the objective of the venture is to define phrases properly. This could be accomplished by translating the punctuation marks such as a period mark.", "inside the text analyzed. The next step is to procure the words (tokens) from phrases. In this phase, the task is to manage the transcription in one phrase. Specifying mistakes, URLs and punctuation from the subsequent token collection are therefore to be rectified. As seen in the figure below after a tweet is tokenized, the returned result is a table containing a set of strings

- ***Stemming and Lemmatization***

The objective of both stemming and lemmatization is to minimize fusional modes and notations of a word to a prevalent basic form, For example, the preceding words: "construction", "constructs", "constructive", "constructed", "constructing" has the same stronghold, which is "construct". Stemming is indeed a primitive heuristic mechanism that really peels off the finishes of words so that only the base form is kept. By contradistinction, lemmatization utilizes the morphological observation of the words, reverting their thesaurus form (core), commonly referred to as the lemma. However, this method depends on a dictionary for a language such as English instead of a language that is more morphologically diversified. In fact, a lemmatizer may incorporate uncertainty by recommending all feasible lemmas for a word form, or by picking the wrong proposal from two conflicting lemmas (e.g., ax the plural of an ax or of an axis) It comprises of five stages, where word cuts are conducted.

- ***Part of Speech Tagging (POS)***

The connection between its phrases must be created in an attempt to comprehend the full significance of a phrase. This could be undertaken by entrusting each term a classification that designates the lexical features of that term.

Also recognized as part of speech tagging (POS), this phase can be seen as a supplementary prerequisite for n-grams availability and lemmatization.

➢ ***N-grams***

N-gram is an ordinary text mining method, where term sub-sets of entire length n are created within a phrase. From the phrase " This is a phrase of six phrases! "Could be created from the preceding N-grams

As either, the instance sentence above might generate 6 unigrams, 5 bigrams, and 4 trigrams. On even a larger data set, generating bigrams and trigrams will considerably make a contribution to the shape of the data set, subsequently, trying to slow down the program. A strategy to this challenge

The preprocessing results will be consistent and consistent information that can be used to maximize the performance of the classifier. Once the various preprocessing measures are implemented, we can now concentrate on the learning portion of the machine.

F. Feature Extraction and Sentiment Classification

After the text is divided into sentences, the words were tokenized and normalized, each sentence being segmented into words. We can build a simple text model "bag of words.".

a. Bag of words

In this bag-of-words representation, you only take individual words into account and give each word a specific subjectivity score. This scoring of subjectivity can be seen in a lexicon of feelings. If the overall result is negative, the document is categorized as adverse and the message is favorable.

The lexicon of emotion can be developed with a few easy practice statistics of the training set to do it the class probability of each word portray in the bag-of-words will be calculated

The sentiment lexicon is incredibly easy to make, but it is less precise since it does not take the sentence structure of the grammar into account. The use of bigrams and trigrams is a straightforward enhancement. This is to avoid dividing a sentence by the words' not," no," very," just' etc. It is simple to use but can improve precision significantly.

The precise words to put in a bag-of-words include important words that give loosely coupled information and unfair words that assist to clear differentiation of polarities (Mulcrone, 2012). The bag-of-words model merely uses a statistical method to identify polarities.

b. *Text blob*

The Text blob package for Python is a popular way of doing a lot of Natural Language Processing (NLP) functions. For instance, from text blob import Text Blob Textblob("not a very great calculation) "sentiment. This informs us that, with the English word "not a very large calculation," it is about-0,3 polarity. This means it is somewhat adverse and about 0.6 subjectivity. Such helpful remarks give us more data about the figures we want: Each term in the lexicon has a label and therefore each word in the lexicon has scores.

IV. MACHINE LEARNING CLASSIFICATION

The phrase machine learning refers to the "automatic identification of significant trends in information" Along with the increasing size of the data produced, machine learning is becoming a popular tactic for knowledge extraction. Machine training is implemented in a wide range of fields, ranging from spam filters, voice recognition, tweet removal and personal commercials to google campaigns and image detecting technology.

- The rest of this section will present machine learning algorithms used to classify the polarity (positive, negative, neutral) of the tweets in their normalized form. While the variety of present algorithms based on the learning task, specialized literature makes the distinction according to the nature of the interaction between the computer and the environment. As such, the separation is made between supervised algorithms and unsupervised algorithms of machine learning

A. Supervised Learning

Several techniques of supervised classification will be explored in this framework and which is used to predict the sentiment class. This method for classification includes two sets: firstly, a sequence of training sets used to train the classifier in order to understand how to modify sentence and text characteristics. Furthermore, the test data is used to verify the classification algorithm output. The guided training technique, such as Naive Bayes, was proven more successful in the identification of sentiment classification or text classification

B. Un Supervised learning

Unsupervised machine learning algorithms have the same applicability as supervised learning, which would map an input to output. However, the key distinction is that in the training phase the input is not classified, subsequently, the machine has to find a framework in the input, without specifically being told how to identify. A monitored strategy was appropriate as a portion of the venture.

C. Machine learning classification in brief

Several supervised classification techniques will be explored in this framework. And what would be used to estimate the sentiment class In order to learn to modify the features of a word or text, this classification algorithm involves two sets: first, a series of trainings used to train the classification. Furthermore, test data are used to verify the classification algorithm efficiency. The supervised form of master training, such as Naïve Bayes, was checked as most successful in identification of emotions.

1. Naive Bayes Classification

The Naïve Bayes Classifier is a Probabilistic Classifier, a sub-classification of the Supervised Learning Method. It utilizes combination models, which strongly suspect each class is an element of the system.

NBC is perhaps the most renowned probabilistic classification and it was apparent that it would be almost generic when researching its operation. This classification is among the most consistently used classification, mainly because of its simplicity, due to its simple mathematics.

The model operates with a bag of words, i.e a collection of words that are unordered. Each word is maintained in its frequency, but not in position. Vectorization is used to convert tweets into numerical measurements. Three primary measures are involved in vectorization: tweet tokenization, counting, and normalization. The Bayes' theorem depicts the probability of the phenomenon of an activity based on an associate event regarding the same theory, the Naïve Bayes Classifier tends to take the judgments of determining the class of a model. In our implementation, the characteristics of each tweet are obtained using the module learn. feature extraction module from the class sklearn of Scikit-learn. Many language processing tasks are classification tasks, although our classes are fortunately much easier to define. We introduce a description of naive Bayes algorithms showing the mission of classifying a full text by assigning a letter tag taken from certain sets of labeling to the significant classification problem of text categorization.

Naïve Bayes algorithm is perhaps the most extensively used and it is a simple yet effective supervised classification technique. The basic theory of the technique is to measure the probabilities of sentiment (either positive or negative) for the given perception using the mutual probabilities of a set of words in a specific category. This method is totally reliant on the naïve hypothesis of the term independence. Naive Bayes works fast for the training phase.

Naïve Bayes is one of the most strengthened classifications (classifier) techniques. First, in order to achieve classification, we must customize the features from the data set. All the tweets in the dataset will be extracted by the classifiers. A classifier for the Naive Bayes implies that there is nothing to do with the existence of a specific function in a category. A naïve Bayes algorithm is very easy to construct and predominantly used

i. Conditional Probability

The algorithm is based on the principle of conditional probability. Conditional probability is a measurement of the probability of a specified event. In Bayes theorem, we conclude that events are conditionally independent.

The formula preceding is used to determine the probability of condition:

$$P(A \mid B) = P(A \cap B)/P(B)$$

In which L.H.S symbolizes conditional probability of A Given B, although R.H.S is a quotient of the probability of the combination of events A and B, and the probability of

ii. Prior Probability:

The prior probability of an occurrence is the probability of the occurrence computed before the series of new data. A prior probability is indeed the probability of how an assessment will come down into a set before you gather the data. For illustration, if you are categorizing the buyers of a particular vehicle, you may already understand that 60% of purchasers are male and 40% are female. If you understand or can calculate these probabilities, discriminant analysis can use other prior probabilities in evaluating the posterior probabilities.

iii. Likelihood:

The likelihood is the sample size which is the probability of certain measured results in terms of provided set of variable values which are is considered as the likelihood of the collection of parameter values by the observed outcomes.

iv. Posterior Probability:

The posterior probability of phase A is the probability of phase B occurring. The subsequent probability of a spontaneous scenario or an unclear proposal is the dependent probability that is allocated concerning the appropriate evidence or background.

v. Bayes Theorem

Bayes theorem offers a way of analyzing posterior probability from prior probability and likelihood. Suppose A is the main event and B is the dark cloud event. Bayes theorem then computes the probability as:

$$P(A \mid B) = P(A)\, P(B \mid A) / P(B)$$

Here the argument is made that occurrences are conditionally independent. Here P(A) symbolizes posterior probability, P(A) and P(B) are prior probabilities of A and B sequentially and P(B|A) is the likelihood.

C. How Naïve Bayes Algorithm Works for Sentiment Analysis

As an example, let us try and find the probability that a tweet (the document) can be classified as positive (the class). Because the probability of tweet P(tweet) is constant, our calculations can be overlooked. We are only interested in the probability of the tweet according to the class, P(tweet/positive) and class probability,

i. P(positive):

P(positive/tweet) =P(tweet/positive)*P(positive)P(positive)

For the sake of this example, let's say there are two possible classes: positive, negative that gives any tweet an in two (or 50%) chance of falling into any of those classes. That gives us P(positive) = 0.50

ii. P(tweet/positive)

To evaluate P(tweet/positive), we need to have a training set of tweets that have already been categorized into two sections. This provides us with a baseline for calculating how likely a tweet will drop into a particular category. As the likelihood of finding a particular tweet in the session are comparatively small, we will tokenize the tweet and calculate the likelihood of each phrase in the workout collection. The aforementioned formula is given below

$$P(\text{tweet/positive}) = P(T1|\text{positive}) * P(T2|\text{positive}) * .. * P(Tn/\text{positive})$$

Where T1 to Tn is all the words in the tweet.

iii. P(T1|positive)

- To determine the probability of a specific word falling into the category we're testing, we'll need the following from the training set:

- The number of times T1 occurs in tweets that were marked as positive in the training set. The total number of words of tweets that were marked as positive in the training set.
- There are various ways in which you can get these numbers, so we won't go into specifics here. As an example, let's look at the word "food", with the following numbers:
  •Number of times food occurs in positive tweets: 455
  •Number of words in positive tweets: 1211
- So to calculate the relative probability of food occurring in the positive category, we divide 455 by 1211, giving us 0.376. Since food can have positive, negative and neutral interpretations, it's not surprising that its relative probability is 37%. This process now needs to be repeated for each word in the tweet.
- Since we now have the ability to calculate the probabilities that each word in the tweet can be classified as positive, let's calculate the probability that the whole tweet can be classified as

  Positive – P(positive/tweet) = P(tweet/positive)* P(positive).
- For this example, let's say the tweet was "I love good food", and the probabilities we calculated were 25%, 62.5%, 74%, and 42.5% respectively.

P(positive/tweet=P(tweet/positive)*P(positive)= P(T1|positive)**P(Tn/positive)*P(positive)=0.25*0.625*0.74*0.425*0.50= 0.024570

- This same procedure can now be used to calculate the relative probability for each of the classes.

  - If P(positive/tweets)>P(negative/tweets) = comment is positive
  - If P(negative/tweets)>P(positive/tweets) = comment is negative

➢ ***Advantages Of Using Naïve Bayes***

Estimating the class of the data set is very easy and quick. This is often used mainly for forecasts for other classes. If the independence presumption continues, the Naive Bayes classification system is more effective, especially compared with other models such as the logistic regression.

➢ ***Disadvantages Of Using Naïve Bayes***

If a categorical variable has an unknown class in a test data set then the system gives a probability of 0 (zero) and cannot make a prediction. This is often referred to as , zero frequency field.' We can use the smoothing method to solve this problem. Laplace estimation is one of the most basic smoothing techniques. Most scientists have categorized emotions using the classifier Naive Bayes. But the classification of Naive Bayes has the greatest drawback that the actual data cannot always fulfill. This therefore influences the accuracy of Naive Bayes classifier

➢ ***Applications of Naive Bayes***

- *Real-time Prediction*:

  The Naive Bayes algorithm is indeed a high speed learning algorithm. Therefore, it can be used for making a prediction in real-time.

- *Multi-class Prediction*:

  This classifier is also well recognized for multi-class prediction function. Again we can predict the statistical likelihood of different classes also.

- *Text classification/ Spam Filtering/ Sentiment Analysis*:

  Naive Bayes classifiers are widely used in the text classification because they have a better outcome in multi-class problems and a higher success rate than other algorithms. It is also used for the detection of spam emails. The main application is a sentimental analysis that predicts whether or not a user wants a certain resource

V. PROPOSED ARCHITECTURE

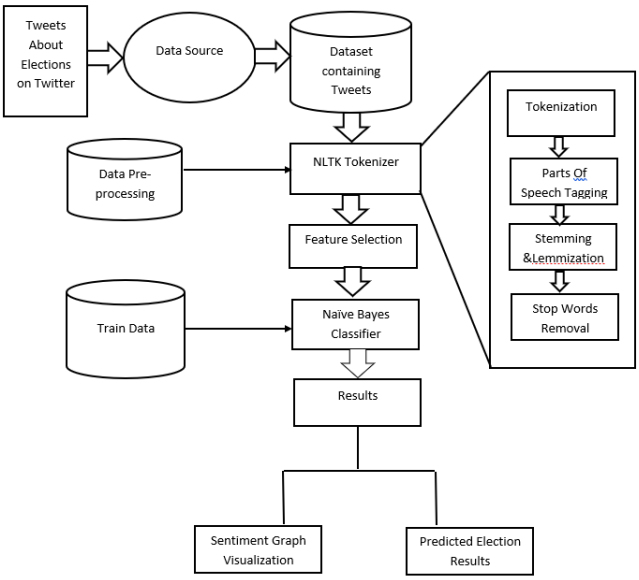

Fig: Proposed Architecture for predicting Election Results

Now we have a dataset, with a two-step marking of this dataset, which can be used to train a supervised machine learning model to assess public sentiment and predict the election results. In order to prepare the train data and test data we split the data set into 80:20 ratios

A. Implementation

The system will deliver a variety of features which include-sentiment analysis of the tweets for the prediction of election results. The first step involves the creation of a dataset that will take tweets as input then extracted in a .csv file. Further, the next step will be preprocessing of tweets that involve following operations- tokenization, stop word removal, stemming and POS (part-of-speech) tagging.

Tokenization is splitting the text into words and then discarding the non-relevant words like pronouns, prepositions, and articles.

There are certain words in the English language such as “I”, “it”, “the”, “of”, which do not carry any meaning. Stop word removal is the process of removing these words.

Stemming is the process in which the slang words and the words which are synonyms will be replaced by their root meaning words. After performing POS tagging, feature selection is done to reduce the amount of data to be investigated and also to identify relevant features for the consideration in the classification process.

This data will be fed to the Naïve Bayes classifier. The Naïve Bayes classifier will classify the given set of words into three different categories- positive, neutral or negative and also, it will find out their corresponding polarity.

The overview of the system is as shown in Fig. Naïve Bayes algorithm is a supervised learning technique. After performing the sentiment analysis, a visual sentiment graph will be generated as the final output and also the polarity will be displayed in the output after considering the text valence in the Naïve Bayes classifier.

Algorithm: *Opinion Mining of predicting Twitter Data :*

Input: collection of all data obtained D

Output: Divided data P //polarized data

1. Initialize Data extracted set D
2. Initialize Selected Token set T

//Switching to Lower case

3. for each t€ D do
4. k ← tweet;
5. if T(k) = NULL then
6. T(k) = p;
7. else T(k)=lowercase ();

//Remove URL

8. for each p € D do
9. k← p. tweet;
10. if T(k)=NO URL then
11. T(k)=p;
12. else

T(k)=p.sub('((www\.[^\s]+)|(https?://[^\s]+))','URL',tweet);

//Removing username

13. foreach p € D do
14. k← p.tweet;
15. T(k) = p.sub('@[^\s]+','AT_USER',tweet);

//eliminating additional white spaces

16. foreach p € D do
17. k← p.tweet;
18. T(k) = p.sub('[\s]+', ' ', tweet);

//Topic classification

19. T = p.sub(‘#word related to the list’,’’)

// Load the subject wise excluded tweets in different data repository

20. foreach p € D do
21. k← p.tweet;
22. T(k)=p.store();

//Polarity Claissifier

23. if(tweet having positive text) then
24. p.positivesentiment();
25. elseif(tweet having negative text) then
26. p.negative sentiment();
27. elseif(tweet having negation) then
28. if(next 3 words are polar noun, verb or adj)
29. p.reversepolarity():
30. elseif(emoticon=TRUE) then
31. if(emoticon=positive) then
32. p.positivesentiment()
33. elseif(emoticon=negative) then
34. p.negativesentiment();
35. else p.neutralsentiment();

VI. EXPERIMENTAL RESULTS

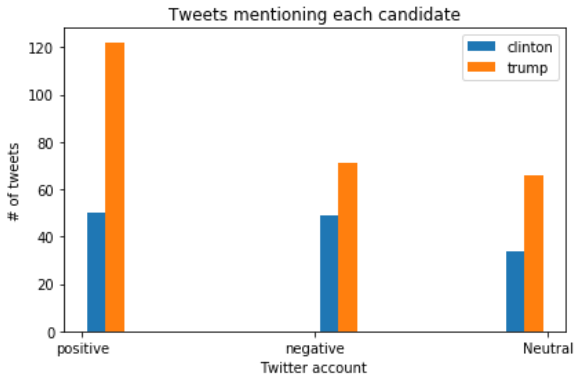

Fig: Tweets Mentioning Each Candidate

In this work, a new method for evaluating sentiments at elections is proposed, based on twitter data sets, tweets are sample public views and this sample tweets may bring us examples of positive and negative sentiments from the judgment of the US government regarding the outcome of the elections

Our initial goal was to develop a method that yielded an accurate sentiment analysis classification in prediction of election results using Twitter as the source of content. During the elections we gathered tweets and filtered these tweets using the Natural Language Toolkit (NTLK),here unnecessary common words are eliminated Such processed tweets were preserved in text files and imported to evaluate the sentiments to assess the viewpoint of the user about the overall contextual sentiment.

1. Hash Tags

In addition to the short messages the user may categorize such tweets to make them appear more quickly on Twitter Search with the hashtag symbol "#" before the appropriate keywords. The use of hashtags improves the issue of text scoring, An emotional opinion can be generated by the hashtag itself. For example, Donald Trump's official slogan is MakeAmericaGreatagain. i.e,All the tweets in this hashtag indicates support for the candidate

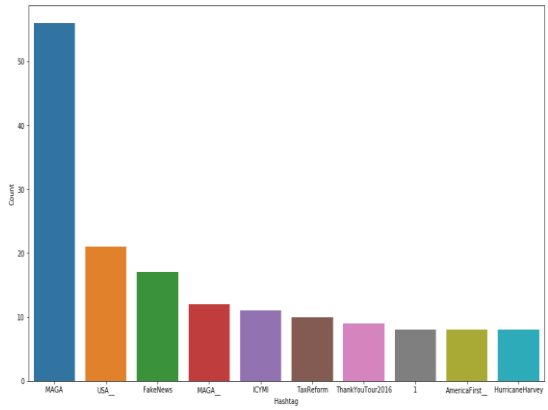

Fig: Hashtags related to trump

We mainly concentrated efforts on classifying positive, negative or neutral. For our proposed model, we perform multistage classification and identify whether the sentiment of a tweet is positive or negative w.r.t. one of the election candidates.

When the tweet was posted about the elections Twitter users mainly mentioned Donald Trump. As a result, when finding messages about Donald Trump, we find many more Tweets than when we looked for Hillary Clinton, utilizing hashtags such as #DonaldTrump, #maga or the negative #NeverTrump. Furthermore, we could say that there was much more people talking about DonaldTrump than Hillary Clinton. It probably doesn't matter that a large proportion of these tweets were negative for Trump. There seems to be no such phenomenon as negative publicity

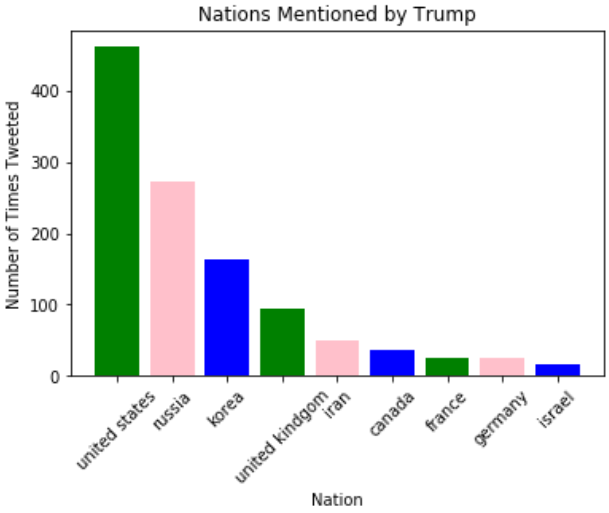

Fig: Nations Mentioned by Trump

Therefore, calculating the percentage of supportive tweets for each politician is going to give a fair idea of each candidate's popularity

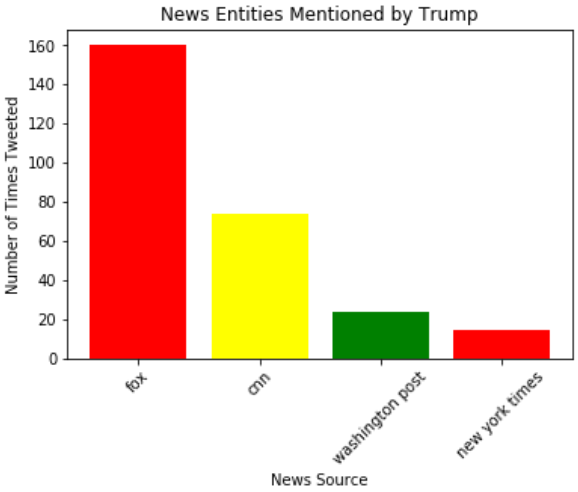

Fig: News entities mentioned by trump

Using Plotly, Word Cloud offers a visual overview of the most common words used in the tweets. The common word occurs prominently in the Word Cloud. Fig: represents the word cloud for which the stored tweets trump record is embedded. Comparably, Hilary word clouds can be collected as well.

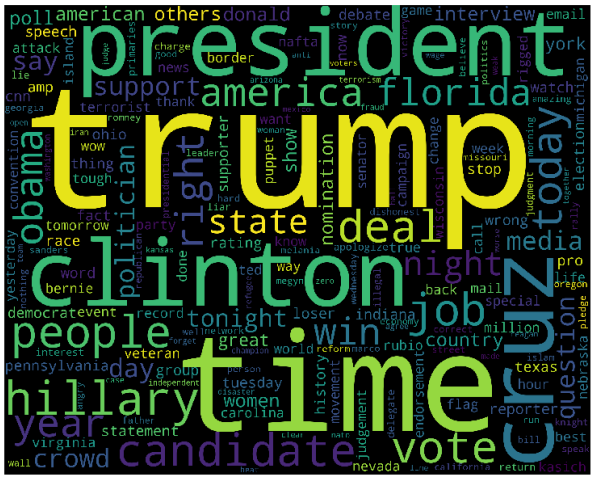

Fig: Word Cloud for trump tweets

VII. CONCLUSION

The objective of this work is to analyze the capability of Twitter data as a measure of Election outcomes. This research analyzes Twitter data related to the US elections in 2016. Currently, the system only focuses on text classification. This research highlighted the importance of identifying certain emotions in a tweet about elections and showed how these specific feelings can be derived using sentiment analysis methods focused on the supervised approach of machine learning i.e. Naive Bayes. This approach is focused on the predicting sentiment of voters. Prior to the announcement of the Indian Election Commission the forecasts of political elections, results were published in social media. It indicates that social media outlets can forecast poll results correctly according to polling results. A brief political analysis is also conducted

Therefore, this justification suggests that the proposed framework will be employed to help accurately predict outcomes by next generation prediction. Therefore, we can conclude that Twitter is a social media platform that forecasters can use to obtain meaningful findings on elections.

REFERENCES

1. Godbole, Namrata, Manja Srinivasaiah, and Steven Skiena. "Large-Scale Sentiment Analysis for News and Blogs." ICWSM 7.21 (2007): 219-222.
2. Fumagalli, Laura.and Emanuela Sala. 2011. "The total survey error paradigm and pre-election polls: the case of the 2006 Italian general elections". Tech. rep., ISER Working Paper Series: 2011-29.
3. Mondher Bouazizi, Tomoaki Ohtsuki, "A Pattern-Based Approach for Multi-Class Sentiment Analysis in Twitter", Access IEEE, vol. 5, pp. 20617-20639, 2017, ISSN 2169-3536.
4. Boia, Marina, et al. "A:) is worth a thousand words: How people attach sentiment to emoticons and words in tweets." Social computing (social com),2013 international conference on. IEEE, 2013.
5. Manuel, K., Kishore Varma Indukuri, and P. Radha Krishna. "Analyzing internet slang for sentiment mining." 2010 Second Vaagdevi International Conference on Information Technology for Real-World Problems. 2010.
6. Akcora, Cuneyt Gurcan, et al. "Identifying breakpoints in public opinion." Proceedings of the first workshop on social media analytics. ACM,2010.
7. Gao, Wei, and Fabrizio Sebastiani. "Tweet sentiment: From classification to quantification." Advances in Social Networks Analysis and Mining(ASONAM), 2015 IEEE/ACM International Conference on. IEEE, 2015.
8. Bhoir, Purtata, and Shilpa Kolte. "Sentiment analysis of movie reviews using lexicon approach." Computational Intelligence and Computing Research (ICCIC), 2015 IEEE International Conference on. IEEE, 2015.
9. Mandal, Santanu, and Sumit Gupta. "A Lexicon-based text classification model to analyze and predict sentiments from online reviews." Computer, Electrical & Communication Engineering (ICCECE), 2016 International Conference on. IEEE, 2016
10. Aung, Khin Zezawar, and Nyein Nyein Myo. "Sentiment analysis of students' comments using a lexicon-based approach." Computer and Information Science (ICIS), 2017 IEEE/ACIS 16th International Conference on. IEEE,2017.
11. Hailong, Zhang, Gan Wenyan, and Jiang Bo. "Machine learning and lexicon-based methods for sentiment classification: A survey." Web Information System and Application Conference (WISA), 2014 11th. IEEE,2014.
12. Rezapour, Rezvaneh, et al. "Identifying the Overlap between Election Result and Candidates' Ranking Based on Hashtag-Enhanced, Lexicon-BasedSentiment Analysis." Semantic Computing (ICSC), 2017 IEEE 11th International Conference on. IEEE, 2017.
13. Ramteke, Jyoti, et al."Election result prediction using Twitter sentiment analysis." Inventive Computation Technologies (ICICT), International Conference on. Vol. 1. IEEE, 2016.
14. Bouazizi, Mondher, and Tomoaki Ohtsuki. "Sentiment analysis in twitter: From classification to the quantification of sentiments within tweets." Global Communications Conference (GLOBECOM), 2016 IEEE. IEEE, 2016.
15. J. M. Soler, F. Cuartero, and M. Roblizo, ``Twitter as a tool for predicting elections results," in Proc. IEEE/ACM Int. Conf. Adv. Social Netw.Anal. Mining (ASONAM), Aug. 2012, pp. 1194_1200.
16. A. Java, X. Song, T. Finin, and B. Tseng, ``Why we Twitter: Understanding microblogging usage and communities," in Proc. 9th WebKDD 1st SNA-KDD Workshop Web Mining Social Netw. Anal., Aug. 2007, pp.56_65. Proceedings of the second International Conference on Inventive Systems and Control (ICISC 2018)IEEE Xplore CompliantPartNumber:CFP18J06-ART,ISBN:978-1-5386-0807-4;DVDPart Number:CFP18J06DVD, ISBN:978-1-5386-0806-7978--
17. Harshali P. Patil,Dr. Mohammad Atique, "Sentiment Analysis for Social Media: A Survey", Department of Computer Engineering Thakur College of Engineering & Technology Mumbai, India.
18. Megha Rathi, Aditya Malik, Daksh Varshney, Rachita Sharma, Sarthak Mendiratta, "Sentiment Analysis of Tweets using Machine Learning Approach", Jaypee Institute of Information Technology JIIT Sec-62 Noida, India.
19. http://www.nltk.org/
20. https://pypi.python.org/pypi/textblob
21. Mondher Bouazizi, Tomoaki Ohtsuki, "A Pattern-Based Approach for Multi-Class Sentiment Analysis in Twitter", Access IEEE, vol. 5, pp. 20617- 20639, 2017, ISSN 2169-3536.

AUTHORS PROFILE

Kambhampati Kalyana Kameswari, M. tech Student, Department of CSE, SRKR Engineering College affiliated to JNTU Kakinada, Bhimavaram, AP,India. Email: kalyani95.kambhampati@gmail.com

J Raghaveni, Assistant Professor of Computer Science and Engineering, SRKR Engineering College affiliated to JNTU Kakinada, Bhimavaram,AP, India. Email: hariveni9@gmail.com

R. Shiva Shankar, Assistant Professor of Computer Science and Engineering, SRKR Engineering College affiliated to JNTU Kakinada, Bhimavaram,AP, India. Email: shiva.srkr@gmail.com

Ch. Someswara Rao, Assistant Professor of Computer Science and Engineering, SRKR Engineering College affiliated to JNTU Kakinada, AP, Bhimavaram, India. Email: chinta.someswararao@gmail.com