e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-10, Issue-2 (April 2020)
https://doi.org/10.31033/ijemr.10.2.12

TWEEZER – Tweets Analysis

Anannya Gupta[1], Deepali[2], Manvesh Ahlawat[3], Vedant[4] and Swati Sharma[5]
[1]B.Tech Student, Department of I.T., MIET, Meerut, INDIA
[2]B.Tech Student, Department of I.T., MIET, Meerut, INDIA
[3]B.Tech Student, Department of I.T., MIET, Meerut, INDIA
[4]B.Tech Student, Department of I.T., MIET, Meerut, INDIA
[5]Assistant Professor, Department of I.T., MIET, Meerut, INDIA

[1]Corresponding Author: anannya.gupta.it.2016@miet.ac.in

ABSTRACT

Twitter is one in all the foremost used applications by the people to precise their opinion and show their sentiments towards different occasions. Sentiment analysis is an approach to retrieve the sentiment through the tweets of the general public. Twitter sentiment analysis is application for sentiment analysis of information which are extracted from the twitter(tweets). With the assistance of twitter people get opinion about several things round the nation .Twitter is one such online social networking website where people post their views regarding to trending topics .It s huge platform having over 317 million users registered from everywhere the globe. a decent sentimental analysis of information of this huge platform can result in achieve many new applications like – Movie reviews, Product reviews, Spam detection, Knowing consumer needs, etc. during this paper, we used two specific algorithm –Naïve Bayes Classifier Algorithm for polarity Classification & Hashtag classification for top modeling. this system individually has some limitations for Sentiment analysis. The goal of this report is to relinquish an introduction to the present fascinating problem and to present a framework which is able to perform sentiment analysis on online mobile reviews by associating modified naïve bayes means algorithm with Naïve bayes classification.

Keywords— Sentiment Analysis, Naïve Bayes, Hashtag Classification, Classification Technique

I. INTRODUCTION

Sentiment Analysis and Opinion Mining consists study of sentiments, attitudes, reactions, evaluation of the content of the text. Twitter may be a micro blogging media in real time to precise the perception of someone or group of couple a particular topic to look occuring a timeline. The message which is displayed on Twitter is called as Tweet. The chronologically sorted collection of multiple tweets is that the the timeline. A person can express his view ahead of globe in front of the world in various forms like multimedia, text etc. Due to popularity of Twitter as an information source, it led to development of applications and research in many spheres. Twitter is employed in predicting the happenings of earthquakes and identifying relevant users to follow to obtain disaster relevant information. Web search applications, Real world applications like world events, current trending topics in world, extracting latest information about incidents uses by the micro blog data for their analysis and conclusion making on the particular topics.

A good sentimental analysis of knowledge of this huge platform can results in achieve many new applications like – Movie reviews, Product reviews, Spam detection, Knowing consumer needs, etc. During this paper, we used two specific algorithm -Naïve Bayes Classifier Algorithm for polarity Classification & Hashtag classification for top modeling. This technique individually has some limitations for Sentiment analysis.

II. LITERATURE SURVEY

In most of the conventional literature on sentiment analysis, researchers have addressed the binary task of separating text into Positive and Negative categories

[1]. However, there is early work on building classifiers for first detecting if a text is Subjective or Objective by separating Subjective text into Positive and Negative classes

[2]. The definition of Subjective class contains only Positive and Negative classes, in contrast to more recent work of Wilson et al who additionally consider Neutral class to be part of Subjective class. It build classifiers for the binary task Subjective versus Objective or the ternary task Neutral, Positive and Negative. However, they do not explore the 4-way design and the cascaded design. One of the earliest work to explore these design issues

[3]. They compare a 3-way classifier that separates news snippets into one of three categories: Neutral, Positive and Negative, to a cascaded design of two classifiers: Polar versus Non-polar and Positive versus Negative. They defined Polar to contain both Positive and Negative class or Non-polar to contain only Neutral class. We extend on their work to compare a 4-way classifier to a cascaded design of three models are Objective versus Subjective, Polar versus Non-polar and Positive versus Negative. This extension poses a question about training the Polar versus Non-polar model: should Non-polar category contain Neutral examples or both Neutral and Objective. Of course, the 4-way classifier puts all three categories Objective, Positive and Negative are together while training a model to detect Neutral. In this paper, we explore these designs. In the circumstance of micro-blogs such as Twitter, to the best of our knowledge, we know of no literature that explores this issue. It build two separate classifiers, one for Subjective versus Objective classes and one for Positive versus Negative classes

[4]. They present separate evaluation on both models but do not explore combining them or comparing it with a 3- way classification. More recently, [3] present results on building a 3-way classifier for Objective, Positive and Negative tweets. Yet, they do not explore the cascaded design and do not detect Neutral tweets. Moreover, to the best of our paradigm, there is no work in the literature that studies the trade-off between making less predictions and F1-measure. Like human annotations, postulation made by machines have confidence levels. In this paper, we compare the 3 classifier designs in terms of their ability to predict better given a chance to make predictions only on examples they are most confident on.

III. PROBLEM STATEMENT

A major advantage of social media is that we are able to see the good and bad things people say about the actual brand or personality. The larger your company gets difficult it becomes to stay a handle on how everyone feels about your brand. For large companies with thousands of daily mentions on social media, news sites and blogs, it's extremely difficult to do this manually. To combat this problem, sentimental analysis software is necessary. This software's can be used to evaluate the people's sentiment about particular brand or personality.

IV. PROPOSED METHODOLOGY

1. Retrieval of Tweets

As twitter is the most enlarged part of social networking site, it consists of various blogs which are related to various topics in the world. Instead of taking whole blogs, we will rather search on particular topic and download all its pages then extracted them in the form of text files by using mining tool i.e. Weka which provides sentiment classifier..

2. Pre-processing of Removed Data

After retrieval of tweets Sentiment analysis tool is applied on untested tweets but in most of cases results to very poor performance. Therefore ,preprocessing techniques are necessary for obtaining better results as given. We extract tweets i.e. short messages from twitter which are used as untested data. This data needs to be pre-processed. So, pre-processing involves following steps which constructs n-grams:

i) Filtering:

Filtering is nothing but extraction of raw data. In this step, URL links (E.g. http://twitter.com), special words in twitter, user names in twitter e.g. @Ron - @ symbol indicating a user name, emoticons are extracted.

ii) Tokenization:

Tokenization is nothing but partitioning of sentences. In this step, we will tokenize or segment text with the help of partitioning text by spaces and punctuation marks to form container of words.

iii) Removal of Stopwords:

Articles like "a", "an", "the" and other stopwords such as "to", "of", "is", "are", "this", "for" removed in this step.

iv) Construction of n-Grams:

n-grams can make out of consecutive words. Negation words such as "no", "not" is attached to a word which follows and precedes it. For Instance: "I do not like remix music" has two bigrams: I do+not, do+not like, not+like remix sentence, So the correctness of the classification improves by such procedure, because negation plays an important role in sentiment analysis.

Paper [3] represents that negation needs to be taken into account, because it is a very common linguistic construction that affects polarity.

3. Parallel Processing:

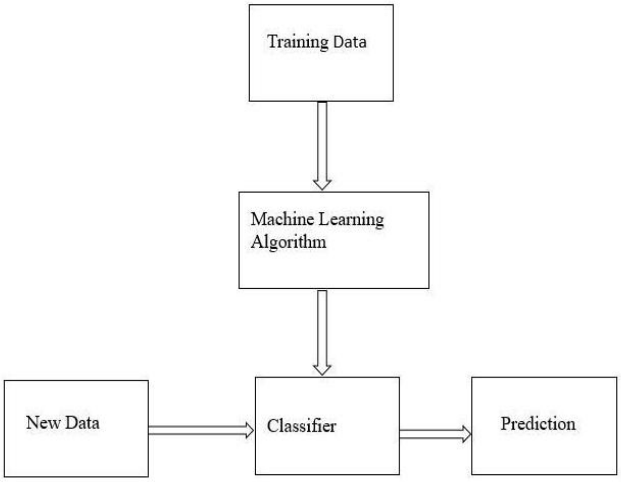
Sentiment classifier which differentiate the sentiments builds using multinomial Naïve Bayes Classifier. Training of classifier data is the main purpose of this module. Every database has hidden information which can be used for decision and prediction are two forms of data analysis which can be used to extract models describing important data and future tendency. Classification is process of finding a set of models or functions that describe and differentiate data concepts, for the purpose of being able to use the model for predicting the class of objects whose class label is not known.

V. SENTIMENTAL ANALYSIS

Sentiment analysis deals with identifying and classifying opinions or sentiments which are present in source text. Social media is generating a larger amount of sentiment rich data within the type of tweets. Sentiment analysis of this user generated data is very useful in knowing the opinion of the mass. Sentiment analysis task is very much fielded specific. Tweets are classified as positive, negative and neutral based on the sentiment present. Out of the entire tweets are examined by humans and annotated as 1 for Positive, 0 for Neutral and a couple of Negative emotions. For classification of non human annotated tweets, a machine learning model is trained whose features are extracted from the human annotated tweets.

Tweets

The word 'micro' in micro blogging specifies the limitation of content of the opinion expressed on that. A twitter user can compose at max 140 characters per each tweet. A tweet isn't only a straightforward text but it is a mix of text data and Meta data associated with the tweet. These attributes are the features of tweets. They expresses the content of the tweet or what's that tweet about. The Metadata are often used to seek out the domain of the tweet. The Metadata of tweet are some entities and places. These entities include user mentions, hashtags, URLs, and media.



VI. PROPOSED MODEL

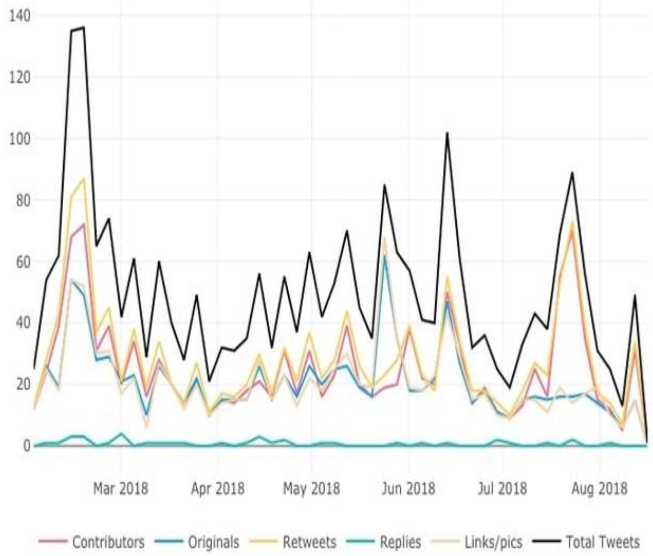
Proposed architecture for sentiment classification. The system deals with the tweets extraction and sentiment classification. It consists of following modules.

A) **DATA COLLECTION**: Accessing tweets from Twitter is the primary requirement for building a dataset to get processed and extract information. Twitter allows its users to retrieve real time streaming tweets by using twitter API. We propose to use the python library Tweepy which has the API to extract the tweets through authenticating connection with Twitter server. While collecting tweets we filter out retweets.

B) DATA PREPROCESSING - The data extracted from twitter contains lot of special characters and unnecessary data which we not require. If data is not processed beforehand, it could affect the accuracy as well as performance of the network down the lane. So it is very important to process this data before training. Extracting the sentiment from a tweet is not a huge matter as the data found on micro blogging websites contains slang, abbreviations and Twitter specific symbols. The processed tweet requires to be cleared from URL, @ mentions and other Twitter specific symbols such as '#' whilst maintaining the text of the hashtag as it can contain an important reference to the sentiment of the tweet.

C) TRAIN CLASSIFIER - To train the classifier model we will be using a labelled dataset in which every single tweet is labelled as positive or negative based on sentiment.

D) DATA VISUALIZATION - The final step of this process is to take in the classified tweets and generate pie chart to visualize the results. The most frequent words in the dataset can be used to generate word cloud.



E) SENTIMENTAL ANALYSIS OF TWEETS – Sentiment Analysis is classification of the polarity of a given text in the document, sentence or phrase. The goal is to determine whether the expressed opinion in the text is positive, negative or neutral.

Tokenization: Tokenization is the process by which big quantity of text is divided into smaller parts called tokens.

Cleaning Data: By removing the numbers, punctuations, Lowercases, Part of speech tagging. **Remove Stop Words**: One of the major forms of pre-processing is to filter out useless data. In natural language processing, useless words (data), are referred to as stop words.

VII. CLASSIFICATION

Rule-based systems that perform sentiment analysis supported a collection of manually crafted rules.
Automatic systems that rely on machine learning techniques to be told from data.

Hybrid systems that combine both rule based and automatic approaches.

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-10, Issue-2 (April 2020)
https://doi.org/10.31033/ijemr.10.2.12

Machine Learning

Machine learning (**ML**) is the study of computer algorithms that improve automatically through experience. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so .Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or infeasible to develop conventional algorithms to perform the needed tasks.

Machine learning is closely related to computational statistics, which focuses on making predictions using computers.

In its application across business problems, machine learning is also referred to as predictive analytics.

Bayes Theorem

Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:

$$P(y|x) = \frac{p(x|y)\,p(y)}{P(x)}$$

where X- Tuples, y-Hypothesis, P(y|X) represents Posterior probability of y conditioned on X i.e. Probability that Hypothesis holds true given the value of X, P(y) represents Prior probability of y i.e the Probability that H holds true irrespective of the tuple values, P(X|y) represents posterior probability of X conditioned on y i.e. the Probability that X will have certain values for a given Hypothesis, P(X) represents Prior probability of X. The proposed system understands whether tweet is positive or negative with dictionary method. The formula is given

$$\textbf{ACCURACY} = \frac{\sum\textbf{TRUE POS}+\sum\textbf{TRUE NEG}}{\textbf{TOTAL NO OF WORDS}}$$

Natural Language Processing

Natural language processing (NLP) is a field of artificial intelligence in which computers analyze, understand, and derive meaning from human language in a smart and useful way. By utilizing NLP, developers can organize and structure knowledge to perform tasks such as automatic summarization, translation, named entity recognition, relationship extraction, sentiment analysis, speech recognition, and topic segmentation.

"Apart from common word processor operations that treat text like a mere sequence of symbols, NLP considers the hierarchical structure of language: several words make a phrase, several phrases make a sentence and, ultimately, sentences convey ideas," "By analyzing language for its meaning, NLP systems have long filled useful roles, such as correcting grammar, converting speech to text and automatically translating between languages."

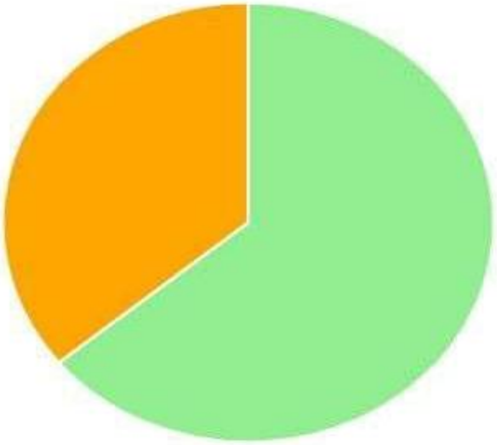
VIII. HASHTAGS CLASSIFICATION

Hashtag classification is very important for topic modeling. While posting any message, the user uses a hash tag, for eg. #Covid19. So, from this we can know that the post is about the Corona Virus 19. This can help in classifying the pre-processed data in various topics. We do not change the hash tag words during pre-processing. With the help of data parsing, the algorithm can identify the hash tagged words and with the help of that particular text message is classified into that group so that the data does not get mixed up and because of that accuracy increases.

In general, people write hash tags in a concatenated format. There are no white spaces or special character in between which parser can identify to split the text. For example Corona Virus 19, tweets related to all covid19 had a hash tag '#Covid19' or **'#CoronavirusTruth' or '#CoronascaresDelhi'**.

IX. RESULT

We will obtain a classification of polarities of sentiments into positive, negative or neutral. Naïve Bayes is easy, easy to coach and has less execution time it shows the output within the kind of pie charts. Thus the fundamental knowledge required to do sentiment analysis of Twitter. Two methods are used in this project. The accuracy/ result of each method enable us to imagine the efficiency of applied technique in respective circumstances.



Figurer: Pie-Chart Representation

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-10, Issue-2 (April 2020)
https://doi.org/10.31033/ijemr.10.2.12

Tweezer

Discover the Twitter sentiment for a product or brand.

iphone 7 plus

IX. CONCLUSION

Nowadays, sentiment analysis or opinion mining is a hot topic in machine learning. We are still far to detect the sentiments of s corpus of texts very accurately because of the complexity in the English language and even more if we consider other languages such as Chinese. In this project we tried to show the basic way of classifying tweets into positive or negative category using Naive Bayes as baseline and how language models are related to the Naive Bayes and can produce better results. We could further improve our classifier by trying to extract more features from the tweets, trying different kinds of features, tuning the parameters of the naïve Bayes.

REFERENCES

[1] Ramteke, Jyoti, Samarth Shah, Darshan Godhia, & Aadil Shaikh. (2016). Election result prediction using Twitter sentiment analysis. In *International Conference on Inventive Computation Technologies (ICICT)*, *1*, pp. 1-5.

[2] Sanjay Kalamdhad, Shivendra Dubey, & Mukesh M. (2016). Feature based sentiment analysis of product reviews using modified PMI-IR method. *International Journal of Computer Trends and Technology*, *34*(2), 115-121..

[3] Sunil Kumar Khatri, Himanshu Singhal, & Prashant Johri. (2017). *Sentimental analysis to predict Bombay stock exchange using artificial neural network*. Available at: https://www.semanticscholar.org/paper/Sentiment-analysis-to-predict-Bombay-stock-exchange-Khatri-Singhal/76aafe74ad8d3228830ec11f114f10df42b82c81.

[4] Bouazizi, Mondher & Tomoaki Ohtsuki. (2016). Sentiment analysis in twitter: From classification to quantification of sentiments within tweets. In *IEEE Global Communications Conference (GLOBECOM*), pp. 1-6.

[5] Wang, Hao, Dogan Can, Abe Kazemzadeh, François Bar, & Shrikanth Narayanan. (2012). A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*, pp. 115-120.

[6] Sang-Hyun Cho & Hang-Bong Kang. (2016). *Text sentiment classification for SNS-based marketing using domain sentiment dictionary*. Available at: https://www.semanticscholar.org/paper/Text-sentiment-classification-for-SNS-based-using-Cho-Kang/7836413461dc5e07c0ab0265e29a787751fd9935.

[7] Zimbra, David, M. Ghiassi, & Sean Lee. (2018). *Brand-related Twitter sentiment analysis using feature engineering and the dynamic architecture for artificial neural networks*. Available at: https://ieeexplore.ieee.org/abstract/document/7427425.

[8] Suman, D.R & Wenjun, Z. (2015). *Social multimedia signals: A signal processing approach to social network phenonmena*. Available at: https://www.springer.com/gp/book/9783319091167.

[9] Liu, B. (2012). *Sentiment analysis and opinion mining*. Available at: https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf.

[10] Kwak Haewoon, Changhyun Lee, Hosung Park, & Moon S. (2010). *What is Twitter, a social network or a news media?*. Available at: http://www.ambuehler.ethz.ch/CDstore/www2010/www/p591.pdf.