EVENT BASED SENTIMENT ANALYSIS ON TWITTER USING MACHINE LEARNING ALGORITHMS

Ashwini Bagade
Department of Computer Engineering,
Zeal College of Engineering and Research Pune, Maharashtra
bagadeashu77@gmail.com

Piyush Jadhav
Department of Computer Engineering,
Zeal College of Engineering and Research Pune, Maharashtra
piyush.jadhav2804@gmail.com

Omkar Yelpale
Department of Computer Engineering,
Zeal College of Engineering and Research Pune, Maharashtra
omkaryelpale2000@gmail.com

Piyush Kulkarni
Department of Computer Engineering,
Zeal College of Engineering and Researc
Pune, Maharashtra
piyushkulkarni3199@gmail.com

ABSTRACT

This project addresses the problem of sentiment analysis in twitter; that is classifying tweets according to the sentiment expressed in them: positive, negative or neutral. Twitter is an online micro- blogging and social-networking platform which allows users to write short status updates of maximum length 140 characters. It is a rapidly expanding service with over 200 million registered users out of which 100 million are active users and half of them log on twitter on a daily basis – generating nearly 250 million tweets per day. Due to this large amount of usage we hope to achieve a reflection of public sentiment by analysing the sentiments expressed in the tweets. Analysing the public sentiment is important for many applications such as firms trying to find out the response of their products in the market, predicting political elections and predicting socioeconomic phenomena like stock exchange. The aim of this project is to develop a functional classifier for accurate and automatic sentiment classification of an unknown tweet stream for analysing the particular . In this project we have used the python library to stream the live raw data from the twitter and we will combine it with the standardised dataset. Then using some Machine Learning algorithms we will carry out the sentiment analysis. This project targets the tweets about the specific event so it will be helpful to use this project analyse the impact of that event over the social media.

Keywords :
NLP, Machine Learning, Naive Bayes, Precision, Recall,f1-score.

1 INTRODUCTION

Nowadays, the popular social networking sites like Twitter, Face- book, Instagram, YouTube and etc. are in trend. The main aim of present work is Sentiment Analysis to dig up the person's behavior, mood, opinion, experience from text data. More than tons of text data on social sites are not written in proper manner, by collecting information manually from that unstructured data is a very difficult task. The main motivation for choosing this domain for the project is we felt that it is immensely necessary to analyse the behaviour of the people during certain events and how it gets reflected on Social Media. We chose Twitter for the analysis because we can get tons of data about people's tweets in very raw format. By using this system we are intended to analyse the sentiments of a specific person's tweets or the overall tweets that we could analyse during the particular event.

In the recent past India has faced so many problems such as Covid- 19 pandemic, Farmer's protest, Student's protest etc. we can see that those were the topics which were vastly discussed over social media. We felt that if we can analyse these events collectively then it will help in finding the solution for the problem, analysing the feelings of the people etc.

2 RELATED WORK

Liza et al [1] they suggest three phases of text mining i.e. pre-processing, processing and validation. After applying primary pre- processing, it performs weighting schemes and use Naïve Bayes as a classification algorithm. Then after in validation phase uses 10-fold cross validation testing.

Shikha Tiwari et al [2] they suggest analytical model for the dataset.This paper suggests the implementation of the model for document level analysis , sentence level analysis and entiny level analysis. The main algorithms used in this model is SVM and the random forest algorithm.

Vishal A. Kharde and S.S. Sonawane [3] in their paper suggested a technique for comparative study of existing techniques for opinion mining including machine learning and lexicon-based approaches. Bhawani Selvaretnam et al [4] they have published a paper on the use of NLP for the sentiment analysis of the dataset. The methodol- ogy of this paper follows the three steps i.e Subjectivity Association, Semantic Association, Polarity Classification.

Pouria Kaviani and Mrs. Sunita Dhotre [5] have published paper on the Short Survey on Naive Bayes Algorithm and discussed the effects of Naive Bayes algorithm on the classification of the senti- ments.

3 SYSTEM ARCHITECTURE

Here, In this paper we discuss only sentence level analysis in which analysis of emotion is done by each sentence that is how each sentence expresses any types of emotion that are mentioned above. If the sentence's emotion is expressionless, then it falls into the neutral category emotion as similarly it distinguishes the happy, sad and all other emotion from the given text. Figure 1.

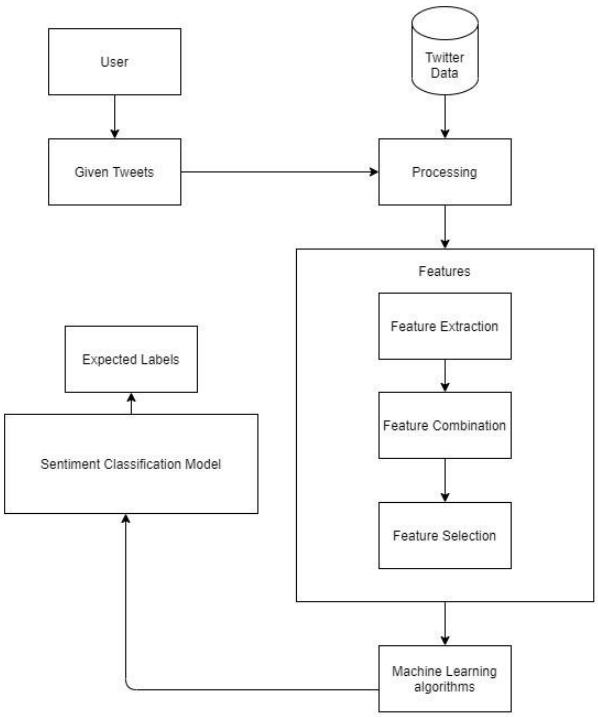

Figure 1: System architecture for proposed System.

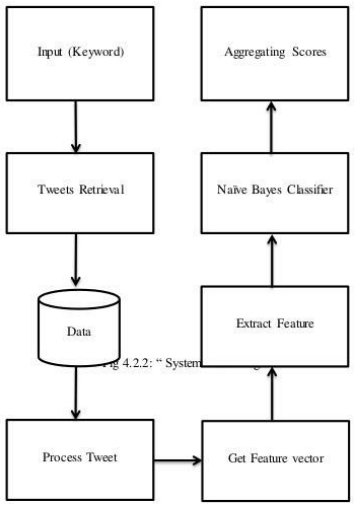

Figure 2: Data flow Diagram

4 METHODOLOGY

In this paper we have came up with the standard model for the processing the data and creating the model.
Following are the crucial steps we followed.

4.1 Sentiment Analysis on Twitter

Sentiment analysis is contextual mining of textual content which identifies and extracts subjective data in supply material, and sup- porting an enterprise to recognize the social sentiment in their brand, services or products at the same time as monitoring on line conversations. Sentiment analysis is also called as subjective analy- sis, it classifies the text according to the polarity and orientation of the opinion expressed into positive, neutral and negative.
It helps facts analysts inside massive organizations gauge public opinion, conduct nuanced marketplace studies, display brand and product popularity, and recognize patron experiences. The process of sentiment analysis consists of Sentiment Identification, Feature Selection, Sentiment Polarity and Sentiment Classification.

4.2 Pre-processing

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

- Remove all URLs (e.g. www.xyz.com), hash tags (e.g. topic), targets (@username)
- Correct the spellings; sequence of repeated characters is to be handled
- Replace all the emoticons with their sentiment.
- Remove all punctuations ,symbols, numbers
- Remove Stop Words
- ExpandAcronyms(wecanuseaacronymdictionary) Remove Non-English Tweet

s

4.3 Feature extraction and selection

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

4.4 Detection process

After completing the feature selection process the next process is to detect the sentiment by using the following attributes.
True Positive: This defines the correct Real Positive Reviews in the testing data
False Positive: It defines the incorrect Fake Positive Reviews in the testing data.
True Negative: This defines the correct Real Negative Reviews in the testing data.
False Negative: It defines the incorrect Fake Negative Reviews in

the testing data.

4.5 Sentiment Classification and result analysis

The final process is Sentiment classification, which is done using machine learning algorithms. Here first the raw data is cleaned and then stored as CSV.Then this CSV is used as the dataset to create the python-based model for further calculations.The confusion ma- trix was also evaluated to test the working of the applied algorithms.

In this paper the following performance evaluations have been introduced. It includes Real Positive Reviews, Fake Positive Reviews, Real Negative Reviews and Fake Negative Reviews. Performance Metrics

(1) F-measure

F-measure means the arithmetic mean of precision and recall. F-measure is used as assessment metric to analyze the views of sentiment classification.we have acheived the of0.7811475 The f1-score can be calculated as follows

$$2 * \frac{Precision * Recall}{Precision + Recall}$$

(2)Precision and recall

Precision and Recall compares the overall performance of tex- tual content mining. Precision is used to evaluate correctness and recall is used to evaluate the completeness. Precision means the proportion of number of examples correctly la- beled as positive to the number of examples classified as positive labels. Recall means the proportion of number of examples correctly labeled as positive to the total number of examples labeled as positive.

The precision is calculated as :

$$\frac{TP}{TP + FP}$$

The Recall is calculated as:

$$\frac{TP}{TP + FN}$$

In this model we achieved the precision value of 0.7811475 and recall of value 0.78114

(3)Confusion matrix

A confusion matrix consists of detailed information about the predicted and actual values. Performance of the classification values is evaluated using the data in the matrix. In confusion matrix the X axis consists of predicted labels and the Y axis consists of true labels.

5 CONCLUSION AND FUTURE WORK

The task of sentiment analysis, especially in the domain of micro- blogging, is still in the developing stage and far from complete. So, we propose a couple of ideas which we feel are worth exploring in the future and may result in further improved performance. Right now, we have worked with only the very simplest unigram models; we can improve those models by adding extra information like closeness of the word with a negation word. We could specify a window prior to the word (a window could for example be of 2 or 3 words) under consideration and the effect of negation may be incorporated into the model if it lies within that window. The closer the negation word is to the unigram word whose prior polarity is to be calculated. Apart from this, we are currently only focusing on unigrams and the effect of bigrams and trigrams may be explored. As reported in the literature review section when bigrams are used along with unigrams this usually enhances performance. However, for bigrams and trigrams to be an effective feature we need a much more labelled data set than our meagre 9,000 tweets. Right now, we are exploring Parts of Speech separate from the unigram models, we can try to incorporate POS information within our unigram models in future. So, say instead of calculating a single probability for each word like P (word | obj) we could instead have multiple probabilities for each according to the Part of Speech the word belongs to. For example we may have P(word | obj, verb), P (word | obj, noun) and P (word | obj, adjective). One more feature we that is worth exploring is whether the information about relative position of word in a tweet has any effect on the performance of the classifier.

In this research we are focussing on general sentiment analysis. There is potential of work in the field of sentiment analysis with par- tially known context. For example, we noticed that users generally use our website for specific types of keywords which can divided into a couple of distinct classes, namely: politics/politicians, celebri- ties, products/brands, sports/sportsmen, media/movies/music. So, we can attempt to perform separate sentiment analysis on tweets that only belong to one of these classes (i.e., the training data would not be general but specific to one of these categories) and com- pare the results we get if we apply general sentiment analysis on it instead.

REFERENCES

[1] “Liza Mikarsa, Sherly Novianti Thahir, "A Text Mining Application of Emotion Classifications of Twitter’s user using Naïve Bayes Method", IEEE, 2015”
[2] Shikha Tiwari,Anshika Verma et al“"Social Media Sentiment Analysis Twitter Datasets "ICACCS 2020”
[3] Vishal A. Kharde ,S.S. Sonawane “Sentiment Analysis of Twitter Data: A Survey of Techniques IJCA 2016”
[4] Bhawani Selvaretnam et al “Natural Language Processing for Sentiment Analysis. International Conference on Artificial Intelligence with Applications in Engineering and Technology 2014”
[5] Pouria Kaviani, Mrs. Sunita Dhotre “Short Survey on Naive Bayes Algorithm. International Journal of Advance Engineering and Research Development 2017”
[6]Radhi D. Desai “Sentiment Analysis of Twitter Data IEEE 2018”
[7]https://www.wikipedia.org/