Sentiment Analysis of Twitter using Machine Learning

Sushmita Hattarkar[1] Prajakta Tamse[2] Kajal Waghmare[3] Sonal Balpande[4]
[1,2,3]Student [4]Assistant Professor
[1,2,3,4]Department of Computer Engineering
[1,2,3,4]K.C College of Engineering & Management Studies & Research (Mumbai University),Thane, Maharashtra, India

Abstract— Sentiment Analysis (SA) is breaking new grounds in the field of data analysis, it has gained the limelight of many researchers, forasmuch as analysis of twitter text is worthy and favorable to the company as well as customers in many different aspects. This paper gives us a glimpse about how tweets can be analyzed and utilized by the organizations with the ability to scrutinize communities' emotion towards the events or products interrelated to them. Sorting out thousands of tweets would be an arduous task for a human to yield a potential result regarding that particular topic. Our objective is based on the approach of classifying tweets into three categories which can be positive negative or neutral. We made certain to use a classification strategy based on Naive Bayes (NB) because it is a facile and intuitive method for analysis, NB combine efficiency with plausible accuracy. The result of the sentiment analysis on twitter data will be displayed in a graph with different sections presenting positive, negative and neutral sentiments. This helped us to bring to a successful conclusion in defining particular tweets and built a review of a particular product.

Key words: Sentimental Analysis: Reviews; Naïve Bayes Algorithm; Twitter

I. INTRODUCTION

With the huge amount of increase in web technologies, the no of people expressing their views and the opinion via the web are increasing. This data is beneficial for everybody like businesses, governments, and individuals .with 500+ million tweets per day, Twitter is thus turning into a significant source of data which can be used by professionals in their field.

Twitter being a microblogging website, which is popularly known for its short messages known as tweets. It has a limit of 140 characters. Twitter features a user base of 240+ million active users and thus it's a helpful supply of data. The users usually discuss their personal views on varied subjects and additionally on current affairs via tweets. Out of all popular social media like Facebook, Twitter, Google+, and Myspace we choose Twitter because of the reasons like tweets are shorter and therefore are less unbiased and Twitter's audience varies from ordinary Twitter users to country's president to celebrities, politicians and various authorities . Therefore it's doable to gather text posts of users from totally different social and interests' teams.

On contrary customers may conjointly find out about positivity or negativity of different features of products/services according to users' opinions, to make an educated purchase. Furthermore, applications like rating movies, hotels, flights based on online reviews could not emerge without making use of corpus like these. The popularity of Sentimental Analysis is increasing among social networking and it is one amongst the foremost mostly studied applications of Natural Language Processing (NLP) and Machine Learning (ML).

II. OBJECTIVE

Most of the data is neither organized nor in pre-defined form, such text is difficult to analyze, comprehend and sort through. Sentiment Analysis thus grants and provides the ability to bring sense to this big sea of data by providing a vision and thus in case saving a lot of time in manual data processing.

- The intention of this project is to demonstrate how Sentiment Analysis will assist an organization or users experience over a Social network.
- A supervised learning algorithm is used to classify the obtained tweets, which will learn about the emotions from statistical data which will then perform Sentiment Analysis
- Maintaining accuracy in predicting in the final output is our main preference.
- How a congregation discerns about a product is the main purpose of Sentiment Analysis.
- The collected data from Twitter will be classified into three categories which will be classified as positive, negative and neutral.
- An analysis will be performed based on Naïve Bayes Algorithm which predicts the result.
- Particular importance is mainly placed on evaluating the supervised machine learning algorithm Naïve Bayes for the task of Twitter sentiment analysis since it helps to figure out precise probabilities

III. LITERATURE SURVEY

With the population of blogs and social networks, opinion mining and sentiment analysis became a field of interest for many researches. A very broad summary of the prevailing work was conferred in (Pang and Lee, 2008). In their survey, the authors describe existing techniques and approaches for an opinion-oriented information retrieval. However, not several researches in opinion mining thought of blogs and even a lot of less self-addressed microblogging.

In (Yang et al., 2007), the authors use web-blogs to construct a corpora for sentiment analysis and use feeling icons appointed to diary posts as indicators of users' mood. The authors applied SVM and CRF learners to classify sentiments at the sentence level so investigated many strategies to see the general sentiment of the document. As the result, the winning strategy is defined by considering the sentiment of the last sentence of the document as the sentiment at the document level

Some of the first and up to date results on sentiment analysis of Twitter information ar by Go et al. (2009), (Bermingham and Smeaton, 2010) and Pak and Paroubek (2010). Go et al. (2009) use distant learning to acquire

sentiment data.They use tweets ending in positive emoticons like ":)" ":-)" as positive and negative emoticons like ":(" ":-(" as negative. They build models like Naive Bayes, MaxEnt and Support Vector Machines (SVM), and they report SVM outperforms other classifiers. In terms of feature house, they try a Unigram, Bigram model in conjunction with parts-of-speech (POS) features. They note that the unigram model outperforms all different models. Specifically, bigrams and POS features do not help.

Pak and Paroubek (2010) collect information following the same distant learning paradigm.

They perform a distinct classification task though: subjective versus objective.

For subjective information they collect the tweets ending with emoticons within the same manner as Go et al. (2009).

For objective information they crawl twitter accounts of in style newspapers like "New York Times", "Washington Posts" etc. They report that POS and bigrams both help (contrary to results presented by Go et al. (2009)).

Both these approaches, however, are primarily based on ngram models. Moreover, the info they use for coaching and testing is collected by search queries and is thus biased.

In contrast, we present features that achieve a significant gain over a unigram baseline.

In addition we tend to explore a distinct methodology of knowledge illustration and report vital improvement over the unigram models.

Another contribution of this paper is that we tend to report results on manually annotated information that doesn't suffer from any notable biases. Our data is a random sample of streaming tweets unlike data collected by using specific queries..

IV. METHODOLOGY

In our project we are classifying Reviews from Twitter, these reviews are obtained from a specific keyword. To explain the project in detail, consider it in different stages. Our analysis of Twitter reviews is depended mainly on a supervised learning algorithm. We have used the Naïve Bayes Algorithm and Support Vector Machine Algorithm. Both the algorithm are best used for sentiment analysis.

A. Twitter API to Fetch Twitter Data

We first developed a Twitter API for downloading the tweets. From there we obtained the Authentication keys and tokens, these keys helps to connect to twitter and fetch all necessary tweets from twitter. In further steps, a keyword is taken from the user, on the basis of that keyword tweets are fetched from the twitter.

B. About GUI

A GUI is displayed which provides the user a search textbox in addition to it the user can also specify the number of tweets he/she wants to fetch. The keyword entered by the user is accepted as a query by a function, the language is already set as English so as to avoid other unnecessary tweets. In the next step, the obtained tweets are cleaned. Since the user can provide any number of tweets which are to be fetched, thus helping in classifying a large number of tweets at a time.

C. Pre Processing Of Obtained Tweets

The tweets or data obtained from twitter consists of various non-sentiment contents which are not at all useful for the analysis. Thus the data obtained needs to be pre-processed for analysis purpose. Tweets are short sentences which contain URL, Hashtags or any kind of links. These unwanted parts need to be cleaned by removing links, special characters using simple regex statements. Both the uncleansed and newly cleaned tweet obtained are displayed in GUI which helps the user to understand the tweets in a better way

Removing URLs: The twitter data obtain contains different types of data. Many users insert links or URLs which are of no significance for sentiment analysis. Thus such URLs need to be removed from the tweet.

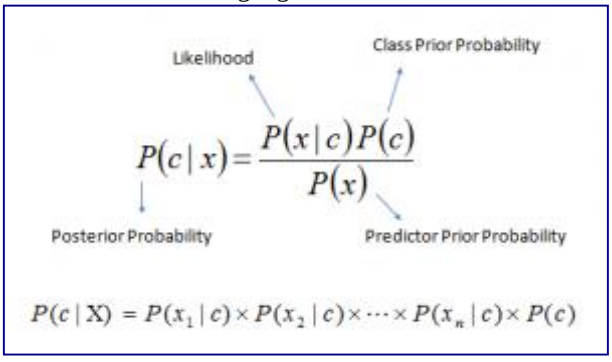
- Removing Special Symbols: Twitter user use various type of symbols Such as period(.), punctuation marks(!),percentage(%),etc. which have zero roles in sentiment analysis. Thus need to be removed.
- Removing Additional White Spaces: Twitter users may enter one or more spaces which are of no use for analysis, removing these white spaces helps in the efficiency of sentiment analysis.
- Removing Hashtag and Username: Usernames are treated as a proper noun and thus need not to be kept in the data, also hashtag has no significant role thus are removed.
- Breaking sentences into Words: The sentences thus obtained are breakdown into a single word, so as to perform sentimental analysis on them.

D. Overview of Tools and Libraries

We have used python 2.7 version and various libraries such as Tweepy, Scikit, Scikit-Learn, Sklearn, Pandas, Textblob, matplotlib, etc

E. Application of Naïve Bayes Algorithm

Naïve Bayes classification is a representation of supervised machine learning method and a statistical method as well for classification. An assurance for capturing uncertainty regarding the model in a moralistic way by determining probabilities is provided by this probabilistic model. With a combination of appropriate past knowledge and observed data, Bayesian classification which provides a useful learning algorithm. It is beneficial in providing a valuable perspective for understanding and evaluating various different learning algorithms.



P(C | X) is posterior probability, P(X | C) is a likelihood, P(C) is class prior probability, P(X) is predictor prior probability. The way of calculating posterior probabily is provided by Bayes Theorem, P(c|x), from P(c), P(x), and P(x|c). This Naïve Bayes classifier is based on the assumption that the effect of the value of predictor(x) on given class (c) is independent of values of other predictors. This presumption is known as conditional independence.

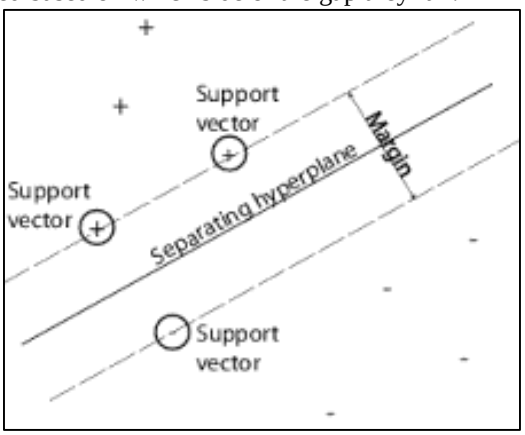
Above,

- P(c|x) is the posterior probability of class (c, target) given that the predictor (x, attributes).
- Prior probability of class is P(c).
- The probability of predictor given class is P(x|c).
- Prior probability of predictor is P(x)

In our project the tweets obtained are accepted as a query, this query is then compared with the words which were used to train the Naïve Bayes algorithm. The algorithm then returns the data into a classified form which can be positive, negative, and neutral. For Example, the user searches for a keyword, then tweets related to that keyword will be fetched and all the tweets will be analyzed for its positive, negative and neutral value. These all will be summed up together and a proper percentage value will be returned. This value will be displayed on the GUI as well as represented through a pie chart.

F. Application of SVM Algorithm

Support vector machine is a supervised learning algorithm which helps in the analysis of data. When SVM is given a set of training data, this data is already marked as belonging to one or the other category for classification. Then the SVM training algorithm learns it and builds a model that assigns examples to one category or another category. The examples of separate categories are mapped so that they are then separated by a clear gap in them which is as wide as possible. When new examples are given as test data set then they are mapped based on which side of the gap they fall.



We first trained SVM by a training set which contains around 6000 words which are classified as positive, negative and neutral. Due to which three categories were generated namely positive, negative and neutral. The new or unknown words when appears in the tweet are compared to the particular category or region and the value is returned on the basis of the minimum distance from particular categories margin. Various libraries like pandas, sci-kit were used. When tested with various twitter input gives a reliable result. The result is shown in the form of a pie graph which shows the percentage of positive, negative and neutral analysis.

V. RESULT

Firstly the data for analysis is obtained from Twitter. This data obtained needs to be pre-processed before evaluation in order to remove unwanted garbage or noise. Algorithms used were Naïve Bayes and SVM, which were trained by training dataset of words which has a combination of all positive, negative and neutral words labeled particularly. This analysis of tweeter tweets are based on the keyword specified by the user, it can be a particular topic or some trending news. The result of the tweets is stored in an excel sheet which can be used for further reference. Both the algorithm performs analysis and classifies the bunch of tweets obtained into how positive, negative and neutral percentage they are and is also represented into a pie chart.

VI. CONCLUSION

This paper has illustrated that effective sentiment analysis can be performed on a tweet by collecting public tweets or opinions from Twitter. Throughout the duration of this project different machine learning algorithms were employed also twitter API was used to collect data, which was then clean and filtered for sentiment analysis. Such an analysis could provide valuable feedback to producers, traders and help them to spot a negative turn in the viewer's perception of their product and take some improvised measures for the same. As the twitter data is large in the form of opinion, reviews, complaint, feedback, analyzing them becomes a bit tedious job. For this, we have made use of supervised learning algorithms. By analyzing people's reactions, views, the traders can make changes for the betterment of their products. It is apparent from this study that the machine learning classifier used has a major effect on the overall accuracy of the analysis.

ACKNOWLEDGEMENT

We would like to thank our guide Assistant Prof. (Mrs.) Sonal Balpande, K. C College of Engineering & Management Studies& Research, Thane for initiating us into this field of research and for providing us with the necessary guidance, great encouragement throughout the preparation of this paper. We record our deep indebtedness to them for their support. We take this opportunity to express our gratitude to the Staff members, Non-Teaching Staff members and Research Scholars of the Department of Computer Engineering, K. C College of Engineering & Management Studies & Research, Thane, for their timely help and encouragement.

REFERENCES

[1] Bhagyashri Wagh, J. V. Shinde, N. R. Wankhade, 'Sentimental Analysis on Twitter Data using Naive Bayes', International Journal of Advanced Research in Computer and Communication Engineering, December 2016.

[2] Huma Parveen, Prof. Shikha Pandey, 'Sentiment Analysis on Twitter Data-set using Naive Bayes Algorithm', 2nd International Conference on Applied

and Theoretical Computing and Communication Technology (iCATccT)2016
[3] Aliza Sarlan1, Chayanit Nadam2, Shuib Basri, ' Twitter Sentiment Analysis 'Computer Information Science Universiti Teknologi PETRONAS Perak, Malaysia
[4] Soumith Chintala, Sentiment Analysis using neural architectures, New York University New York, NY 10012. 2017
[5] Spencer and G. Uchyigit, "Sentiment or: Sentiment Analysis of Twitter Data," Second Joint Conference on Lexicon and Computational Semantics. Brighton: University of Brighton, 2008.
[6] Zhao Jianqiang, Gui Xiaolin, Deep Convolution Neural Networks for Twitter Sentiment Analysis, Xi'an Jiaotong University,(IEEE)2017

