Real Time Twitter Sentiment Analysis using Natural Language Processing

[1]Anupama B S, [2]Rakshith D B, [3]Rahul Kumar M, [4]Navaneeth M
#Department of Computer Science & Engineering,
Siddaganga Institute of Technology,
Tumkur, India

***Abstract*—Social media websites have emerged as one of the platforms to raise users' opinions and influence the way any business is commercialized. Opinion of people matters a lot to analyze how the propagation of information impacts the lives in a large-scale network like Twitter. Data analysis of the tweets determine the polarity and inclination of vast population towards specific topic, item or entity. These days, the applications of such analysis can be easily observed during public elections, movie promotions, brand endorsements and many other fields. In this project, we will go through making a program that analyzes the nature of tweets on a particular topic. The primary aim is to provide a method for analyzing polarity score in noisy twitter streams. This paper reports on the design of a data analysis, extracting vast number of tweets.**

Results classify user's perception via tweets into positive and negative. In this project, we will go through making a program that analyzes the nature of tweets on a particular topic. The user will be able to input a keyword(hashtag) and get the nature on it based on the latest tweets that contain the input keyword. Each tweet extracted classified based on its sentiment whether it is a positive or negative. Data were collected on movie reviews which were on IMDB Website. Naïve Bayes machine learning algorithm was used. The result from this model was tested using various testing metrics. Moreover, our model demonstrates strong performance on mining texts extracted directly from Twitter.

Keywords—Twitter,Natural language processing,Naive Bayes,sentiment analysis, microblogging.

I. INTRODUCTION

As the internet is growing larger, its reach to the masses is becoming wider. Social Media and Microblogging platforms like Twitter, Facebook, Tumblr dominate in spreading encapsulated news and trending topics across the globe at a rapid pace. A topic or news becomes trending if many users are contributing their opinion and judgments, thereby making it a valuable source of online perception on that particular topic. These topics generally intended to spread awareness or to promote political campaigns, public figures during elections, product endorsements, and entertainment like award shows, movies. Large organizations and firms take advantage of people's feedback on these platforms to improve their products and services which further help in enhancing marketing strategies. One such example can be leaking the pictures of the upcoming iPhone to create a hype to extract people's emotions and market the product before its release. Thus, there is a huge potential of discovering and analyzing interesting patterns from the infinite social media data for business-driven applications. Sentiment analysis is the prediction of emotions in a word, sentence, or corpus of documents. It is intended to serve as an application to understand the opinion, attitudes, and emotions expressed within an online mention. The intention is to gain or access an overview of the wider public opinion behind certain topics. Precisely, it is a paradigm of categorizing conversations into positive, negative, or neutral labels. Many people use social media sites for networking with other people and to stay up-to-date with news and current events. These sites (Twitter, Facebook, Instagram, google+) offer a platform for people to voice their opinions. For example, people quickly post their reviews online as soon as they watch a movie and then start a series of comments to discuss the acting skills depicted in the movie. This kind of information forms a basis for people to evaluate, a rate about the performance of not only any movie but about other products and to know about whether it will be a success or not. This type of vast information on these sites can be used for marketing and social studies. Therefore, sentiment analysis has wide applications and includes emotion mining, polarity, classification, and influence analysis. Twitter is an online networking site driven by tweets which are 280 characters limited messages. Thus, the character limit enforces the use of hashtags for text classification. Currently, around 6500 tweets are published per second, which results in approximately 561.6 million tweets per day. These streams of tweets are generally noisy reflecting multi-topic, changing attitudes information is an unfiltered and unstructured format. Twitter sentiment analysis involves the use of natural language processing to extract, identify to characterize the sentiment content.

A. Objective:

• To implement an algorithm for automatic classification of text into positive or negative.

• To determine the opinion of mass in positive or negative towards the subject of interest.

• Graphical representation of the analysis in the form of bar graph, pie-chart.

II. LITERATURE SURVEY

Sentiment analysis is a growing area of Natural Language Processing with research ranging from document level classification to learning the polarity of words and phrases. Given the character limitations on tweets, classifying the sentiment of Twitter messages is most similar to sentence level sentiment analysis however, the informal and specialized language used in tweets, as well as the very nature of the microblogging domain make Twitter sentiment analysis a very different task. It's an open question how well the features and techniques used on more well-formed data will transfer to the microblogging domain. Just in the past year there have been a number of papers looking at Twitter sentiment and buzz Other researchers have begun to explore the use of part-of-speech features but results remain mixed. Features common to microblogging (e.g., emoticons) are also common, but there has been little investigation into the usefulness of existing sentiment resources developed on non-microblogging data. Researchers have also begun to investigate various ways of automatically collecting training data. Several researchers rely on emoticons for defining their training data.Others also use hashtags for creating training data, but they limit their experiments to sentiment/non-sentiment classification, rather than 2-way polarity classification, as we do. We use data mining methods and apply the following Machine Learning algorithm for this second classification to arrive at the best result:

- Naive Bayes :

Naive Bayes Classification: Many language processing tasks are tasks of classification, although luckily our classes are much easier to define than those of Borges. In this classification we present the naive Bayes algorithms classification, demonstrated on an important classification problem: text categorization, the task of classifying an entire text by assigning it a text categorization label drawn from some set of labels. We focus on one common text categorization task, sentiment analysis, the ex-sentiment analysis traction of sentiment, the positive or negative orientation that a writer expresses toward some object.

Naive Bayes classifier is the simplest and the fastest classifier. Many researchers claim to have gotten best results using this classifier. For a given tweet, if we need to find the label for it, we find the probabilities of all the labels, given that feature and then select the label with maximum probability. The accuracy of Unigrams is the lowest at 79.67%. The accuracy increases if we also use Negation detection (81.66%) or higher order n-grams (86.68%). We see that if we use both Negation detection and higher order n-grams, the accuracy is marginally less than just using higher order n-grams (85.92%). We can also note that accuracies for double step classifier are lesser than those for corresponding single step.

- 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," John Rehling, an NLP expert at Meltwater Group, said in *How Natural Language Processing Helps Uncover Social Media Sentiment*. "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. "NLP is used to analyze text, allowing machines to understand how human's speak. This human-computer interaction enables real-world applications like automatic text summarization, sentiment analysis, topic extraction, named entity recognition, parts-of-speech tagging, relationship extraction, stemming, and more.

NLP is commonly used for text mining, machine translation, and automated question answering. NLP is characterized as a difficult problem in computer science. Human language is rarely precise, or plainly spoken. To understand human language is to understand not only the words, but the concepts and how they're linked together to create meaning. Despite language being one of the easiest things for the human mind to learn, the ambiguity of language is what makes natural language processing a difficult problem for computers to master.

Aim of the project

In this project we are going to consider any sports event, movie trending on a given time and analyse the opinion of the public using real time data present on social media platform twitter. By using the tweets, we try to predict or come to a decision based on the mass opinion which are expressed in the tweets. With more than 321 million active users, sending a daily average of 500 million Tweets, Twitter allows businesses to reach a broad audience and connect with customers without intermediaries. On the downside, it's harder for brands to quickly detect negative content, and if it goes viral you might end up with an unexpected PR crisis on your hands. This is one of the reasons why social listening monitoring conversation and feedback in social media has become a crucial process in social media marketing. Monitoring Twitter allows companies to understand their audience, keep on top of what's being said about their brand and their competitors, and discover new trends in the industry. Are users talking positively or negatively about a product? Well, that's exactly what sentiment analysis determines.

2.1.1 Online Commerce

The most general use of sentiment analysis is in ecommerce activities. Websites allows their users to submit their experience about shopping and product qualities. They provide summary for the product and different features of the product by assigning ratings or scores. Customers can easily view opinions and recommendation information on whole product as well as specific product features. Graphical summary of the overall product and its features is presented to users. Popular merchant websites like amazon.com provides review from editors and also from customers with rating information. http://tripadvisor.in is a popular website that provides reviews on hotels, travel destinations. They contain 75 millions opinions and reviews worldwide. Sentiment analysis helps such websites by converting dissatisfied customers into promoters by analyzing this huge volume of opinions.

2.1.2 Voice of the Market (VOM)

Voice of the Market is about determining what customers are feeling about products or services of competitors. Accurate and timely information from the Voice of the Market helps in gaining competitive advantage and new product development.

Detection of such information as early as possible helps in direct and target key marketing campaigns. Sentiment Analysis helps corporate to get customer opinion in realtime. This real-time information helps them to design new marketing strategies, improve product features and can predict chances of product failure. Zhang et al.proposed weakness finder system which can help manufacturers find their product weakness from Chinese reviews by using aspects based sentiment analysis.

2.1.3 Voice of the Customer (VOC)

Voice of the Customer is concern about what individual customer is saying about products or services. It means analyzing the reviews and feedback of the customers. VOC is a key element of Customer Experience Management. VOC helps in identifying new opportunities for product inventions. Extracting customer opinions also helps in identifying functional requirements of the products and some non-functional requirements like performance and cost.

2.1.4 Brand Reputation Management

Brand Reputation Management is concern about managing your reputation in market. Opinions from customers or any other parties can damage or enhance your reputation. Brand Reputation Management (BRM) is a product and company focused rather than customer. Now, one-to-many conversations are taking place online at a high rate. That creates opportunities for organizations to manage and strengthen brand reputation. Now Brand perception is determined not only by advertising, public relations and corporate messaging. Brands are now a sum of the conversations about them. Sentiment analysis helps in determining how company’s brand, product or service is being perceived by community online.

2.1.5 Government

Sentiment analysis helps government in assessing their strength and weaknesses by analyzing opinions from public. For example, “our PM enforced complete nationwide lockdown even when there was no outbreak in our country, kudos to our PM.” this example clearly shows positive sentiment about government. Whether it is tracking citizens’ opinions on a new 108 systems, identifying strengths and weaknesses in a recruitment campaign in government job, assessing success of electronic submission of tax returns, or many other areas, we can see the potential for sentiment analysis.

III. DESIGN AND IMPLEMENTATION

A. *Proposed System*

An idea which can overcome the disadvantages being faced by traditional survey method to get people opinions, to develop a Machine Learning Model by training the model to categorize the tweets based on sentiment of the tweet and make the model as accurate as possible, first the user will give input i.e. the keyword for extracting the tweets and then the extracted tweets will be categorized by the Machine Learning Model which will be either positive or negative tweet and then the output will be displayed in graphical manner for better understanding of the results.

Advantages of Proposed System:

- There is no need to manually start a survey because in twitter there are already available tweets which are opinions of the people
- There is no need to manually take tweets one by one.
- The user just has to download the application. There is no external hardware components required.

Aim of the proposed system:

The application should be functionally competent as such that it must loaded with features that serve the purpose for which it is created. The features in the application should map to the needs of the users which the app is designed to meet. The application UI should be simple enough for a user to understand how the application works. The application should work successfully without crashing, and people all over the world should be able to use the application without any ambiguity.

B. General Working of the System

The application mainly consists of the following tasks

[1] Building and Training the Machine Learning Model

In this step the end user has nothing to do but this step would be done in the background the end user has no knowledge about this process

Fig. 1. Giving the keyword as input

In this step the user has to give the keyword which should be present in all the tweets which we are going to extract from twitter

TABLE I. Preprocess the tweets extracted

In this step also the user has nothing to do but the extracted tweets has to be preprocessed before sending these tweets to the ML model

a. Getting the prediction of sentiment of extracted tweets using our Machine Learning Model

In this step our ML model predicts the sentiment of the tweets but the results will be stored in an array.

- Displaying the results in graphical representation

In this step the results will be displayed to the end user in graphical manner like bar graph or pie chart for better understanding of the results.

C. Activity Diagrams

Another representation of the software is activity diagram which is a graphical representation of workflows of stepwise activities and actions with support for choice, iteration and concurrency. It visually presents a series of actions or flow of control in a system similar to flow chart or a data flow diagram. **Figure 1** shows the step by step procedure of how the Machine Learning Model is built and trained for prediction

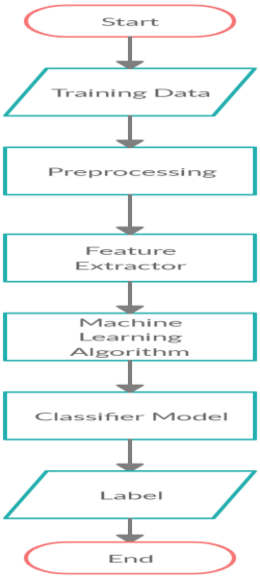


Figure 1: Activity diagram for our ML model

The **Figure 2** shows step by step process of the end user side program

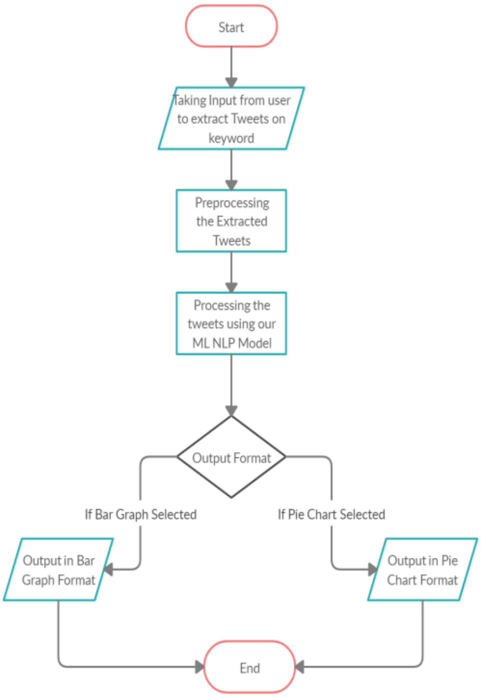

Start
Taking Input from user to extract Tweets on keyword
Preprocessing the Extracted Tweets
Processing the tweets using our ML NLP Model
Output Format
If Bar Graph Selected
If Pie Chart Selected
Output in Bar Graph Format
Output in Pie Chart Format
End

Figure 2: Activity diagram for the user side program

IV. RESULTS

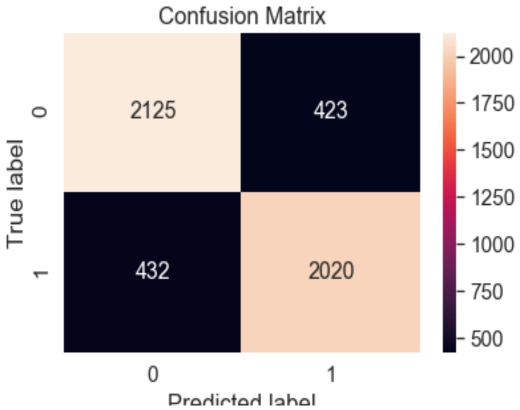
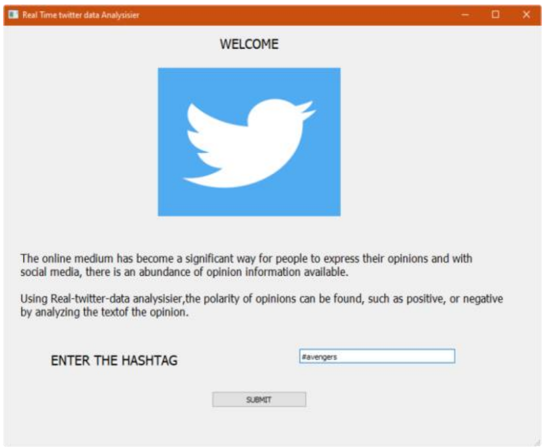


Fig 1: Confusion Matrix of NLP Model(82.9% Accuracy)



The online medium has become a significant way for people to express their opinions and with social media, there is an abundance of opinion information available.

Using Real-twitter-data analysisier,the polarity of opinions can be found, such as positive, or negative by analyzing the textof the opinion.

ENTER THE HASHTAG #avengers

Fig 2: Entry page where we have to enter the hashtag/keyword

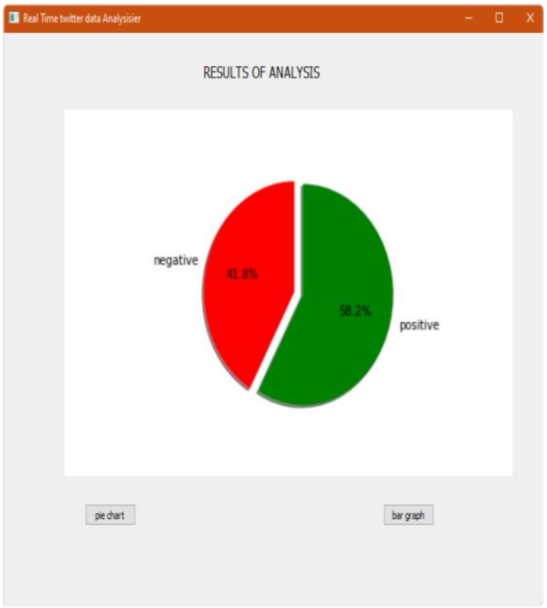


Fig2.1:Results are display in pie-chart

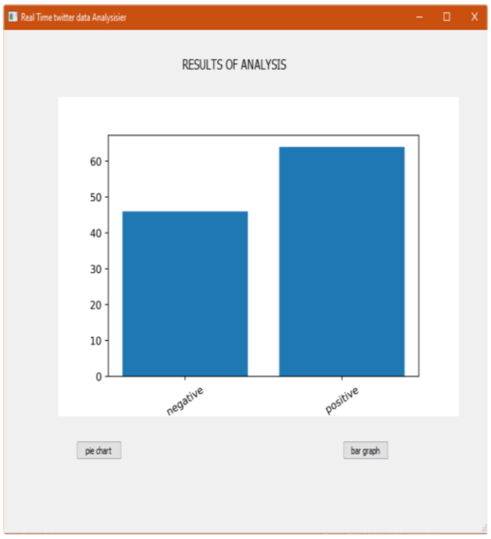
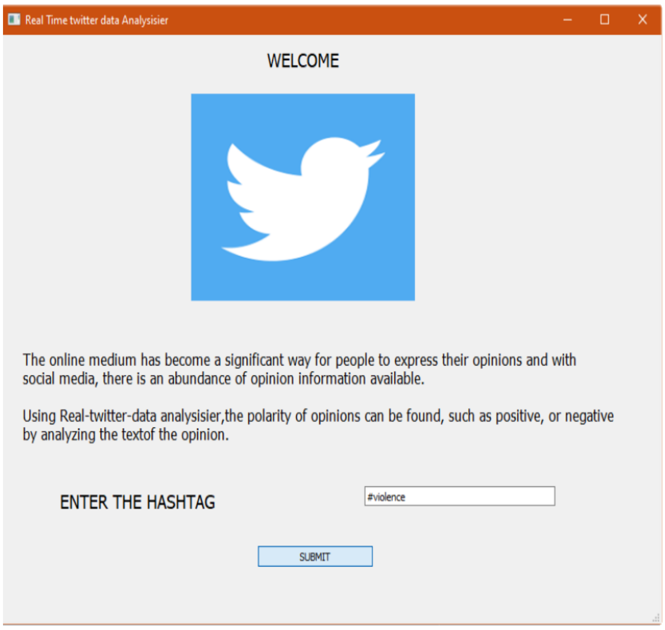


Fig2.2: Results are display in bar graph



The online medium has become a significant way for people to express their opinions and with social media, there is an abundance of opinion information available.

Using Real-twitter-data analysisier,the polarity of opinions can be found, such as positive, or negative by analyzing the textof the opinion.

ENTER THE HASHTAG: #violence

Fig3: Entry page where we have to enter the hashtag/keyword

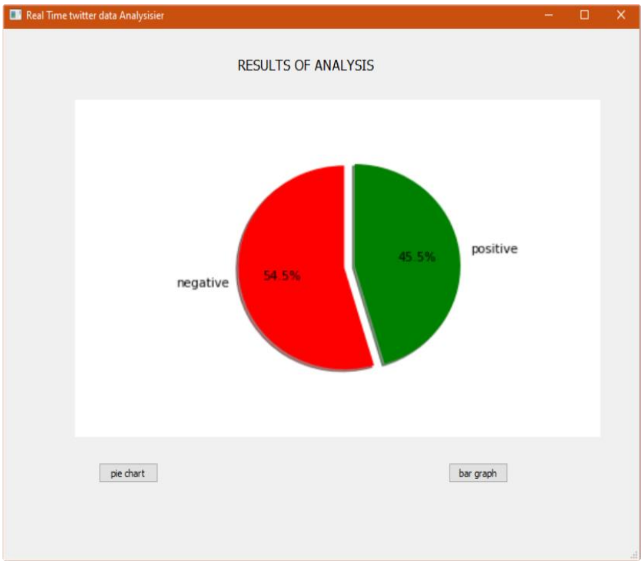


Fig3.1: Results are displayed in pie-chart

V. CONCLUSION

The task of sentiment analysis, especially in the domain of micro-blogging, is still in the developing stage and far from complete. Therefore we would like to propose a couple of ideas which we feel are worth exploring in the future and may result in further improved performance. For now, we have worked with only the very simplest unigram models, we could improve those models by adding additional 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, the more it should affect the polarity. For example, if the negation is right next to the word, it may simply reverse the polarity of that word, and farther the negation is from the word the more minimized ifs effect should be. 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 labeled data set than our meager 9,000 tweets. 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. used a somewhat similar approach and claims that appending POS information for every unigram results in no significant change in performance (with Naive Bayes performing slightly better and SVM having a slight decrease in performance), while there is a significant decrease in accuracy if only adjective unigrams are used as features. However, these results are for the classification of reviews and maybe verified for sentiment analysis on microblogging websites like Twitter. One more feature we that is worth exploring is whether the information about the relative position of the word in a tweet has any effect on the performance of the classifier explored a similar feature and reported negative results, their results were based on reviews which are very different from tweets and they worked on an extremely simple model. In this project, we are focusing on general sentiment analysis. For example, we noticed that users generally use specific types of keywords which can be divided into a couple of distinct classes, namely: media/movies/music, celebrities, products/brands, sports/sportsmen, politics/politicians. 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 compare the results we get if we apply general sentiment analysis on it instead.

FUTURE WORKS

Currently, this project is done using the Naïve Bayes Algorithm which is one of the Machine Learning Algorithm which only got us an accuracy of around 83%. In the future, we will be exploring and implementing the Deep Learning Algorithms to our NLP model in order to increase the accuracy of our model and to get better predictions from our model.

REFERENCES

[1] Efthymios Kouloumpis and Johanna Moore,IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3, July 2012

[2] S. Batra and D. Rao,” Entity Based Sentiment Analysis on Twitter”, Stanford University,2010

[3] Saif M.Mohammad and Xiaodan zhu ,Sentiment Analysis on of social media texts:,2014

[4] Ekaterina kochmar, University of Cambridge, at the Cambridge coding Academy Data Science.2016

[5] Manju Venugopalan and Deepa Gupta, Exploring Sentiment Analysis on Twitter Data, IEEE 2015

[6] Brett Duncan and Yanqing Zhang, Neural Networks for Sentiment Analysis on Twitter.2017

[7] Afroze Ibrahim Baqapuri, Twitter Sentiment Analysis: The Good the Bad and the OMG! Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.2011

[8] Kishori K. Pawar, Pukhraj P Shrishrimal, R. R. Deshmukh,” Twitter Sentiment Analysis: A Review” International Journal of Scientific & Engineering Research, Volume 6, Issue 4, April-2015