Sentiment Analysis of Tweets

Pooja Kumari
Department of Computer Engineering
Padmabhushan Vasantdada Patil Pratishthan's College of Engineering, Eastern Express Highway, Near Everard Nagar, Sion-Chunabhatti, Mumbai-400 022

Shikha Singh
Department of Computer Engineering
Padmabhushan Vasantdada Patil Pratishthan's College of Engineering, Eastern Express Highway, Near Everard Nagar, Sion-Chunabhatti, Mumbai-400 022

Devika More
Department of Computer Engineering
Padmabhushan Vasantdada Patil Pratishthan's College of Engineering, Eastern Express Highway, Near Everard Nagar, Sion-Chunabhatti, Mumbai-400 022

Dakshata Talpade
Department of Computer Engineering
Padmabhushan Vasantdada Patil Pratishthan's College of Engineering, Eastern Express Highway, Near Everard Nagar, Sion-Chunabhatti, Mumbai-400 022

Manjiri Pathak
Department of Computer Engineering
Padmabhushan Vasantdada Patil Pratishthan's College of Engineering, Eastern Express Highway, Near Everard Nagar, Sion-Chunabhatti, Mumbai-400 022

Abstract

Microblogging websites such as twitter have evolved into source of unfettered and wide ranging category of information. Use of socially generated big data to access information about collective states of the minds in human societies becomes a new paradigm in the emerging field of computational social science. One of the natural applications of this would be prediction of the society's reaction to a new product in the sense of popularity and adoption rate. In our paper, we focus on using Twitter, the most popular microblogging platform, for the task of sentiment analysis. Using the corpus, we build a sentiment classifier that is able to determine positive, negative and neutral sentiments for a document.

Keywords: Microblogging websites, Naive Bayes classifier

I. INTRODUCTION

In recent past we noticed an outburst of data availability, the so-called data deluge, determined by an increased amount of social communication performed through different electronic channels. The emergence of internet and web 2.0 led to the outburst of social media providing people an opportunity to publicly share their thoughts and express their opinions. Social media technologies exist in different forms such as blogs, business networks, enterprise social networks, forums, microblogs, photo sharing, products/services review, social bookmarking, social gaming, social networks, video sharing and virtual worlds.

Amongst these, microblogging websites have become a very well-known paradigm for communication. This is due to nature of microblogs on which people post real time messages about their opinions on a variety of topics, discuss current issues, complain, and express positive sentiment for products they use in daily life. Therefore microblogging web-sites have become rich sources of data for opinion mining and sentiment analysis. Sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. The attitude may be his or her judgment or evaluation, affective state (the emotional state of the author while writing), or the intended emotional communication (the emotional effect the author wishes to have on the reader. Along with expressing views and opinions and providing us with novel answers to classic questions, the consummate analysis of this huge amount of data could have practical applications to predict, monitor, and cope with many different type of events, from simple matters of daily life to massive crises in the global scale. At analytical level there are several technological innovations that help making sense of the large amount of data availability. With around 300 million users sending out around 500 millions of micro-blogs (approximately) every day, Twitter is certainly an effective channel for communication. Additionally this social networking site is not just for teenagers or celebrities tweeting about their daily activities but has also emerged as a powerful marketing tool by many business owners who are using it to help their businesses grow. In this paper we will take into account of Twitter, the most popular microblogging platform for the task of sentiment analysis. and build models for classifying "tweets" into positive, negative and neutral sentiment. We build models for two classification tasks: a binary task of classifying sentiment into positive and negative classes and a 3-way task of classifying sentiment into positive, negative and neutral classes.

The remainder of this paper is organized as follows: Section 2 briefly reviews the literature on forecasting the box-office success of theatrical movies. Section 3 gives the details of our methodology by specifically talking about the data, its collection, preprocessing and classification using Naive based classification method. Next, the results evaluated using our method and

experimental results generated by our supervised learning statistical models are shown and explained. Finally, the Section 5 of the paper conclusion that follows and future scope of the topic.

II. LITERATURE SURVEY

In addition to providing us with novel answers to classic questions about individual and social aspects of human life from scientific point of view, precise analysis of this huge amount of data could have practical applications to predict, monitor, and cope with many different type of events, from simple matters of daily life to massive crises in the global scale.

Sentiment analysis has become popular in judging the opinion of consumers over the sentence. In case Twitter[1], the presence of features such as hashtags, emoticons, references have their advantages as well as disadvantages.

Glivia assesses the usefulness of twitter hashtags in sentiment analysis. They analyzed 10,173,382 tweets related to the Brazilian Presidential elections in 2010. They analyzed these tweets and observed that the positive behavior of the tweeters across time was in accordance to the hypothesis that hashtags sentiments match the overall population sentiment. They also verified that the information propagation in twitter follows a cascade model where people make their decisions consciously or not, based on someone else's sentiments and choices.

Statistical analysis of motion picture markets led to intriguing results, such as observing the evidence for a Pareto law for movie income along with a log-normal distribution of the gross income per theater and a bimodal distribution of the number of theaters in which a movie is shown. Previous research in sentiment analysis includes Pang[2] having analyzed the performance of different classifiers on movie review website[3]. The work of Pang acts as a baseline and many authors have used the techniques provided in their paper across different domains. Pang also make use of a similar idea as ours, using star ratings as polarity signals in their training data.

In 2009, Alec Go, Lei Huang, and Richa Bhayani used Twitter to collect training data and then to perform a sentiment search. The approach is similar to Jonathon. The authors construct corpora by using emoticons to obtain "positive" and "negative" samples, and then use various classifiers. The best result was obtained by the Naive Bayes classifier with a mutual information measure for feature selection. The authors were able to obtain up to 81% of accuracy on their test set. However, the method showed a bad performance with three classes ("negative", "positive" and "neutral"). In our work we will classify the data in two classes as positive and negative and include the idea similar as that of Pang to use star signals as polarity rating to summarize the overall data and provide a satisfactory visual output.

III. METHODOLOGY

Our method of sentiment analysis of tweets is based on machine learning technique. Section **i** explains the sources of data we have used. Existing approaches to sentiment analysis can be grouped into four main categories: keyword spotting, lexical affinity, statistical methods, and concept-level techniques. Keyword spotting classifies text by affect categories based on the presence of unambiguous affect words such as happy, sad, afraid and bored. Lexical[4] affinity not only detects obvious affect words, it also assigns arbitrary words a probable "affinity" to particular emotions.

A. Data Collection:

Twitter is a social networking[5] and microblogging service that allows users to post real time messages, called tweets. Tweets are short messages, restricted to 140 characters in length. Due to the nature of this microblogging service (quick and short messages), people use acronyms, make spelling mistakes, use emoticons and other characters that express special meanings. Following is a brief terminology associated with tweets. Emoticons: These are facial expressions pictorially represented using punctuation and letters; they express the user's mood. Target: Users of Twitter use the "@" symbol to refer to other users on the microblog. Referring to other users in this manner automatically alerts them. Hashtags: Users usually use hashtags to mark topics. This is primarily done to increase the visibility of their tweets. For information on how to obtain the data, see Acknowledgments section at the end of the paper. They collected the data by archiving the real-time stream. No language, location or any other kind of restriction was made during the streaming process. In fact, their collection consists of tweets in foreign languages. They use Google translate to convert it into English before the annotation process. Each tweet is labeled by a human annotator as positive, negative, neutral or junk. The "junk" label means that the tweet cannot be understood by a human annotator. A manual analysis of a random sample of tweets labeled as "junk" suggested that many of these tweets were those that were not translated well using Google translate. We eliminate the tweets with junk label for experiments.

B. Pre-Processing (Generation of List of Positive Negative Words):

The pre-processed document is classified with several machine learning algorithms, in this research the algorithms is Naive Bayes to decide best accuracy and performance. Naive Bayes classifier uses a probability to define a document class, the classifier is using statistical approach, even though the classifier is not following the grammatical rules. In Naive Bayes

Classifier the immersion of words does not affect other words immersion, and the absence of a word does not affect the absence of another word, and it does not decrease the accuracy of Naive Bayes Method. Figure 1 shows an example below.

Table 1: Example

Set	Document	Words	Class
Training set	1	I like movie. It's story nice.	pos
	2	Hero's acting is good, I like it but heroin role is bad. Overall movie is fantastic.	pos
	3	I like music, which is so rocking.	pos
	4	Movie story is good but ending is just plain boring and sadly.	neg
Test set	5	I like director's direction. The location place in movie is so boring. But still movie is good.	?

Fig. 1: Example of some positive and some negative emotions.

C. Classification:

Naive Bayes classifier is a simple model for classification. It is simple and works well on text classification. It is a probabilistic classifier based on applying Bayes theorem with strong independence assumptions. This is the simplest form of Bayesian Network, in which all attributes are independent given the value of the class variable. This is called conditional independence. It assumes each feature is conditional independent to other features given the class. A Naive Bayes classifier is a technique that applies to a certain class of problems, namely those that phrased as associating an object with a discrete category. From numerical based approach group, Naive Bayes has several advantages such as simple, fast and high accuracy, in K. Ming Leung describes the Bayes rule as shown in figure 2.

$$\gamma(\alpha \mid \beta) = \frac{\gamma(\alpha) * \gamma(\beta \mid \alpha)}{\gamma(\beta)}$$

Fig. 2: Bayes Rule.

Where α: Specific class, β : Document wants to classify, γ(α) and γ(β) : Prior probabilities γ(α | β) and γ(β | α) : Posterior probabilities. The value of class α might be positive or negative. Document is a review of particular movie. The multinomial model of Naive Bayes captures word frequency information in documents. The Maximum Likelihood Estimate (MLE) is simply the relative frequency and corresponds to the most likely value of each parameter given the training data.

IV. RESULTS

The evaluated results with pie chart, bar graph and star ratings are shown and explained in this section. The movie review dataset is used for this work that is provided. Experimental setup contains simulation environment, parameters and performance metrics. Generally performance metrics are used for calculate one of the metrics like size, execution time, performance accuracy of system. On the basis of this the negative tweets are represented as bar and pie charts below in figure 3.

The result is calculated as by taking input the tweets of many people in different cities and the algorithm decides that either the tweet is negative or positive or neutral. This is respectively shown by pie chart and bar chart in the output screens as shown in the figure.

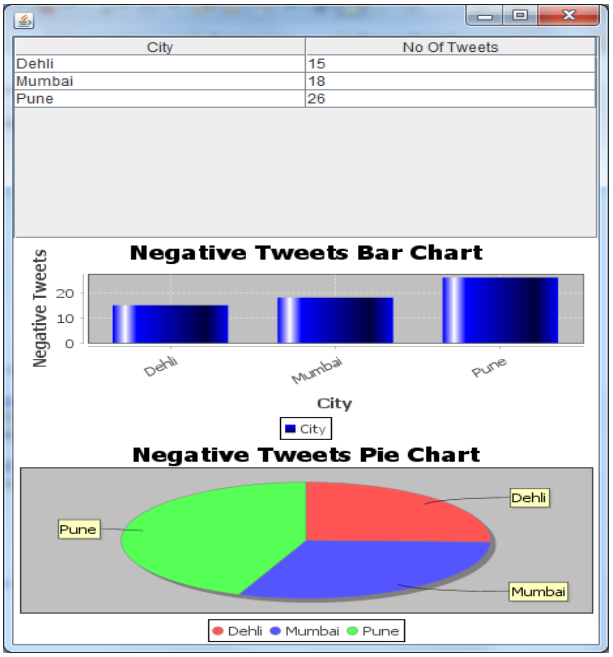

Fig. 3: Negative tweets Bar and Pie Chart.

Similarly the positive tweets are shown in figure 4, where the data read from three cities is represented in bar and pie chart respectively.

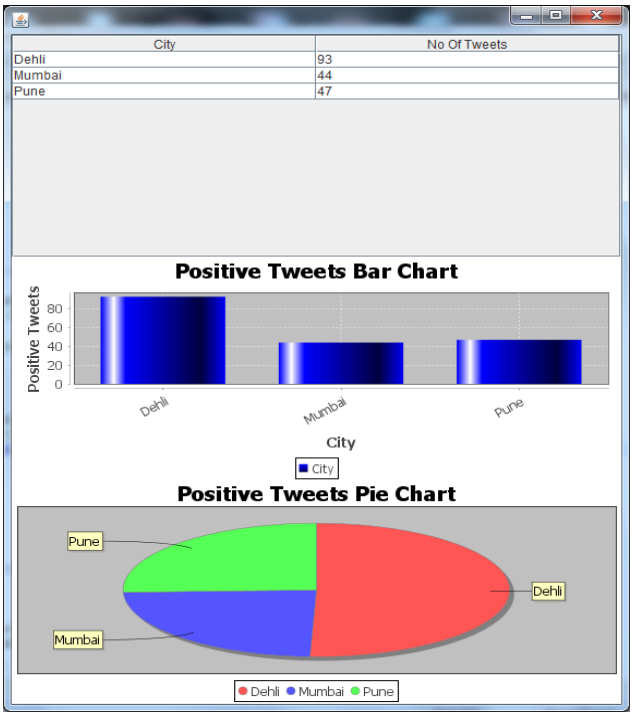

Fig. 4: Positive tweets Bar and Pie chart.

V. CONCLUSION

In this report, we have analyzed the sentiment of microblogs or tweets of social networking and microblogging site, Twitter. We collected the tweets and then classified them using supervised learning algorithm, Naive Bayes classification technique. We further evaluated the polarity of tweets(positive and negative) and presented star ratings based on the calculated polarity. The result shows that this classification method can perform relatively well and can be further refined to determine subjective/objective nature of data. We also observed that this technique can be implemented to any other domain such as product review by altering the list of positive and negative words as per to suit the features of new domain.

VI. FUTURE WORK

The future work in this field includes implementation of this technique on a large scale data set where we have millions of tweets with respect to each domain. Futher improvisation in our work can be done by automated updation of new words in positive and negative keywords list which is done manually. Additionally the sarcasm needs to be detected in the tweets which is not implemented in our work being a very complicated part of natural language[6][7] processing.

REFERENCES

[1] [1] Twitter social networking-"www.twitter.com".
[2] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In In Proceedings of the ACL, pages 271–278, 2004.
[3] Rottentomatoes movie review site. http://www.rottentomatoes.com.
[4] C. Fellbaum. WordNet: An Electronical Lexical Database. The MIT Press, Cambridge, MA, 1998.
[5] Digg social networking site. http://www.digg.com.
[6] Natural language toolkit. http://www.nltk.org.
[7] S. Bird, E. Klein, and E. Loper. Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit. O'Reilly Media, 2009.