ISSN:0975-9646

A Survey on Sentiment Analysis on Twitter Data Using Different Techniques

Bholane Savita Dattu, Prof.Deipali V. Gore

Department of Computer Engineering, Savitribai Phule Pune University
Pune, India

***Abstract*— Sentiment analysis is a broad research area in academic as well as business field. The term sentiment refers to the feelings or opinion of the person towards some particular domain. Hence it is also known as opinion mining. It leads to the subjective impressions towards the domain, not facts. It can be expressed in terms of polarity, reviews or previously by thumbs up and down to denote positive and negative sentiments respectively. Sentiments can be analyzed using NLP, statistics or machine learning techniques. Sentiment analysis may ask questions regarding "customer satisfaction and dissatisfaction, "public opinion towards new iPhone series launched" etc. In real world, public or consumer opinions about some product or brand are very important for its sell. Hence sentiment analysis is a very important research area for real life applications i.e. decision making.**

One of the most visited social networking sites by millions of users is twitter where they share their opinion about various domains like politics, brands, products, celebrities etc.Many research works are carried out in the field of sentiment analysis. But they are only useful in modeling and tracking public sentiments. They had not found exact reasons behind the sentiment variations and hence not useful in decision making.

Sentiment analysis has many applications in various domains like political domain, sociology and real time event detection like earthquakes. Previously research was carried out to model and track public sentiments. But with the advancement in research, today we can use it for interpreting the reasons of the sentiment change in public opinion, mining and summarizing products reviews, to solve the polarity shift problem by performing dual sentiment analysis. Here we use different algorithms/models to perform the above tasks like LDA approach, DSA model, Naïve Bayes (NB) classifier, Support Vector Machine (SVM) algorithm and so on.

Keywords— Opinion mining, Latent Dirichlet Allocation, sentiment analysis, emerging topic mining, event summarization, foreground topics.

I. INTRODUCTION

Sentiment Analysis is to detect the polarity of text in consideration in textual form. It is also known as opinion mining as it derives the opinion of the speaker or the user about some topic. In other words, it determines whether a piece of writing is positive, negative or neutral.

For example, do people on Twitter think that president Barack Obama is doing his job properly or not? To find out the answer we can refer the social networking site twitter. There are millions of opinions of people about Barack Obama, some of them positive and some will be negative or neutral. We can get the exact ideas of why people think Obama is fulfilling his responsibilities or not, by extracting the exact word indicating the positive or negative opinion. It can be carried out at various levels like document level, phrase level or sentence level. When the sentence consists of positive as well as negative sentiments at word level, the whole sentence becomes neutral at sentence level. As the sentiment analysis on twitter or any social media site tracks particular topic, many politicians as well as companies use twitter to track their position in politics and monitor their products and services respectively.

The major benefit of sentiment analysis in previous work was to find out whether the expressed opinion in the document or sentence is positive, negative or neutral. But it was not useful in decision making as no reasons were known about why the sentiments has changed. Hence there was need to build a system for interpreting the public sentiment variations.

Here we have studied different techniques for sentiment analysis like NB classifier, SVM algorithm, NBSVM algorithm etc. for the sentiment analysis. Different researchers have done different work in this domain. They might be real time events like earthquake detection using social sensors, event summarization, interpretation of the public sentiment variations on twitter and so on. These all are the advancements in research as the time goes on. Hence sentiment analysis has become popular field for research work. It is very useful for academic as well as business purposes.

Different Classes of Sentiment Analysis

Sentiments can be classified into three classes .i.e. positive, negative and neutral sentiments.

a. **Positive Sentiments**: These are the good words about the target in consideration. If the positive sentiments are increased, it is referred to be good. In case of product reviews, if the positive reviews about the product are more, it is bought by many customers.

b. **Negative Sentiments**: These are the bad words about the target in consideration. If the negative sentiments are increased, it is discarded from the preference list. In case of product reviews, if the negative reviews about the product are more, no one intend to buy it.

c. **Neutral Sentiments**: These are neither good nor bad words about the target. Hence it is neither preferred nor neglected.

Architectural Diagram for Sentiment Analysis

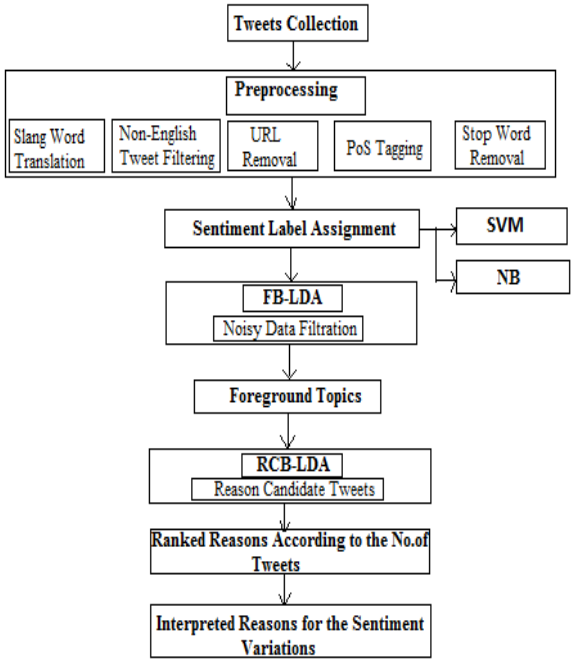

Fig.1.Sentiment Analysis on Twitter Data

Levels of Sentiment classification:

There are three different levels of sentiment classification. i.e. word level, phrase level and document level sentiment classification.

a. **Word Level Classification**: this classification is done on the basis of the words which indicate the sentiment about the target event. The word maybe noun, adjective or adverb. This type of classification gives accurate classified sentiments.

b. **Phrase Level Classification**: This type falls in good as well as bad category. The phrase denoting the opinion is found out from the sentence and the classification is done. But it sometimes gives inaccurate results if a negation word is added in front of the phrase. The phrase refers to combination of two or more words which are closely related to each other.

c. **Document Level Classification**: In this level of classification, single document is considered about the opinionated text. A single review about the single topic from this document is considered. But sometimes it is not beneficial in case of blogs and forums as customers may compare one product with the other which has similar characteristics. Again the document may consist of the irrelevant sentences which don't resemble to opinion about the event.

II. LITERATURE SURVEY

Sentiment analysis is the most important research area in business fields. Previously research was carried out for sentiment analysis in various domains like company product, movie reviews, politics etc. Previous research like Pang et al. has provided with the baseline for carrying out research in various domains. It uses star ratings as polarity signals in their training data. Even many authors have used the same concept provided by Pang et al.

1) *Earthquake shakes twitter users: Real-time event detection by social sensors:*

T.Sakaki et al. [1] developed an event notification system which monitors the tweets and delivers notifications considering the time constraint. They detect real-time events in Twitter such as earthquakes. They have proposed an algorithm to monitor tweets detecting target event. Each Twitter user is considered as a sensor. Kalman filtering and particle filtering are used for estimation of location.

Data set:

For classification of tweets, we prepared 597 positive examples which report earthquake occurrence as a training set.

Advantages:

1. Main task of earthquake detection is done using the system. Users are registered with it and email messages are sent to them.
2. The two filtering techniques detect and provide estimation for location.

Disadvantages:

1. Multiple events cannot be detected at a time.
2. It cannot provide advanced algorithms to expand queries.
3. Limited to only one target event detection at a single time event.
4. It uses SVM as a classifier into positive and negative sentiments which is not applicable to small data sets.

2) *Event summarization using tweets:*

Previous research could not be contributing to detect hidden time events in repeating events such as sports. Here goal is to extract a few tweets that describe the most important stages in that event. Chakrabarti and Punera [2] have described the variation in Hidden Markov Model (HMM) in summarizing the event from tweets. It gives the continuous tweet representation for intermediate stages relevant for an event. Here three algorithms are used to summarize the relevant tweets. HMM gives hidden events.

Data set:

Tweets between the periods Sep 12th, 2010 to Jan 24th, 2011 containing the names of NFL teams.

Advantages:

1. It provides benefits for previous techniques of matching queries.
2. It is most beneficial for one shot events like earthquakes.
3. It tackled problems like construction of real time summaries of events.
4. It identifies an underlying hidden state representation of an event.

Disadvantages:

1. It is not applicable to find continuous time shots present in tweets.
2. It does not give the minimal set of tweets which are relevant to an event.
3. It cannot provide summary of unknown events which cannot be predicted.
4. In this model noises and background topics cannot be eliminated.

3) Et-lda: Joint topic modeling for aligning events and their twitter feedback:

Twitter has become the widely used micro-blogging site to share the opinions. In this work, Y.Hu et.al [3] has proposed a joint Bayesian model ET-LDA that is Event-Topic LDA which performs the task of topic modeling and event segmentation so as to carry out sentiment analysis quantitatively and qualitatively. Here Y.Hu et.al, has taken into consideration two large scale data sets from two different domains associated with two events. The work done here is most useful for topic modeling because the topic may consists of many paragraphs where the tweet may belong to specific event in a paragraph or general event in the topic. So to do the sentiment analysis accurately without misconceptions the event topic model is very useful.

Data Sets:

Two large scale data sets associated with two events from two different domains :(1) President Obama's speech On 19 May 2011 and (2) Republican Primary debate On Sept 7, 2011. Above datasets consist of 25,921 and 121,256 tweets respectively.

Advantages:

1. The baseline LDA treats events and tweets separately while ET-LDA treats them relating to each other. Hence the task of finding polarity and sentiment analysis gives more accurate results.
2. The basic task of event modeling and segmentation of events is carried out successfully.

Disadvantages:

Tweets are modeled as binomial mixture where tweets in which most words belong to general topics are considered as general tweets and tweets in which most words belong to specific event as specific tweets. It is totally unreasonable for tweets having short lengths.

4) An empirical study to address the problem of unbalanced data sets in sentiment classification:

As the web usage has increased all over the globe, sentiment analysis has carried out many researches in academic as well as business fields. But the problem of unbalanced datasets was not solved in these researches. Asmaa M. et al. [4] has addressed the problem using supervised machine learning techniques in multilingual context. The methods to solve the problem are under sampling and over sampling. Here the author finds the under sampling i.e. reduction in the number of documents of the majority class by using the sub-methods like remove similar, remove farthest and remove by clustering. The three classifiers i.e. Support Vector Machine, Naive Bayes and K-NN are used to calculate the accuracy of the sentiments over the three different datasets. Here the Naïve Bayes classifier seems to be insensitive to unbalanced datasets and give more accurate results.

The evaluation measure is g-performance which corresponds to geometric mean of positive and negative accuracies. We use g-performance measure because it is best suited for unbalanced datasets in terms of maximization of the accuracy of the two classes and to balance both the classes at the same time.

Data Sets:

Two Arabic and one English data set are used for the classification. The Arabic datasets are collected from ACOM corpus. It consists of two different domains. First dataset has 468 comments about movie reviews and the second consists of 611 comments about political issues. The English dataset is collected from SINAI corpus which consists of 1846 product reviews.

Advantages:

1. Machine learning methods minimize the structural risks.
2. For prediction of sentiment of documents, supervised machine learning approaches are used.
3. The problem of unbalanced dataset in sentiment classification is solved efficiently and appropriately.
4. Naive bayes classifier seems insensitive to the unbalanced data and gives more accurate results than the support vector machine and K-NN which are sensitive to the unbalanced data. Multilingual sentiment classification is carried out successfully.

Disadvantages:

1. The under sampling method is complex to classify the sentiments and it is a time consuming process.
2. Supervised methods require excessive quantity of labeled training dataset which are very expensive.
3. It may fail when training data are insufficient.

5) Interpreting the Public Sentiment Variations on Twitter:

Twitter sentiment analysis is an important research area for academic as well as business fields for decision making like for the seller to decide if the product should be produced in a large quantity as per the buyers feedback and for the students to decide if the study material to be referred or not. in this work, Shulong Tan et al.[5] have proposed LDA based two models to interpret the sentiment variations on twitter i.e.-LDA to distill out the foreground topics and RCB-LDA to find out the reasons why public sentiments have been changed for the target.

Dataset:

They have considered the twitter dataset for sentiment classification. It is obtained from Stanford Network Analysis Platform. It consists of tweets from June 11, 2009 to December 31, 2009 with 476 million tweets. But the evaluation of results is done on the dataset from June 13, 2009 to October 31, 2009.

Advantages:

1. Distilled out the foreground topics effectively and removed the noisy data accurately.
2. Found the exact reasons behind sentiment variations on twitter data using RCB-LDA model which is very useful for decision making.

Disadvantages:

Uses the sentiment analysis tools TwitterSentiment and SentiStrength whose accuracy is less as compared to other sentiment analysis techniques.

6) *Sentence-Based Sentiment Analysis for Expressive Text-to-Speech:*

Alexander T et al.[6] have proposed a model to tackle the problem of sentence level sentiment classification. They have classified the text into three classes i.e. positive, negative and neutral. The TTS framework is built without using the additional textual data. Till this invention, no attempt was done to use SA methods for TTS requirements. The classifiers are trained to classify the sentiments rely on the representation of the features.

Datasets:

The experiments are performed on two data sets i.e.Semeval 2007 dataset and the twitter dataset. Semeval 2007 dataset consists of news headlines drawn from major newspapers. The corpus has two sets i.e. training data containing 250 headlines and testing data containing 1000 headlines. The twitter dataset consists of tweets with sentences less than 14 words on average.

Advantages:

1. Three class sentiment classification problems at the sentence level have been solved.
2. Additional textual data is not required for classification.i.e.using the unigrams only more accurate and efficient classification results are obtained.

Disadvantages:

1. For the limited size of the training data only the system works properly.
2. The system is applicable only for English tweet classification.

7) *Dual Sentiment Analysis: Considering Two Sides of One Review:*

D. Rui Xia et al. [7] have performed the task of tackling the polarity shift problem. Here the polarity shift causes the negation of the statement. In Bag-of-words technique, two sentiment opposite texts are considered to be very similar which causes the polarity shift. Today most of the researchers use BOW way for sentiment analysis. They have proposed the dual sentiment analysis (DSA) model to solve the polarity shifting. The data is expanded by creating the reversed review for each training and test review. The dual prediction algorithm classifies the test review by considering the two sides of one review. Again they have used DSA3 algorithm to extend the work from polarity classification to the 3-class classification by considering the neutral reviews.

Advantages:

1. Tackled the polarity shift problem in sentiment classification using DSA framework.
2. DSA3 approach is used to extend the work of sentiment classification from polarity shift to 3-class sentiment classification.
3. Corpus based approach is used to construct pseudo antonym dictionary to remove DSAs dependency on an external antonym dictionary for review reversion.

Disadvantages:

Due to the dual nature of each review, the time and space requirement for the classification increases.

III. CLASSIFICATION ALGORITHMS

a) Naïve Bayes Classifier:

The basic mechanism of Naive bayes classifier is done by counting the frequency of words related to sentiment in the message. The tweets are classified and scored according to the number of matches to the sentimental words. The weight of nodes is adjusted according to the importance of tweets and more accurate result of classified sentiments can be generated.

b) Support Vector Machine:

SVM is generally used for text categorization. SVM gives best results than Naive bayes algorithm in case of text categorization. The basic idea is to find the hyperplane which is represented as the vector w which separates document vector in one class from the vectors in other class.

IV. COMPARATIVE RESULTS

Table 1:

Comparative Results for Sentiment Classification Techniques

Model/Algorithm/Tools	Dataset	Accuracy (%)
SVM	Amazon product review data and ChnSentiCorp dataset	89.8
NB	Amazon product review data and ChnSentiCorp dataset	89.4
NBSVM	SEMEVAL 2013 twitter dataset	79.4
MNB	SEMEVAL 2007 twitter dataset	71.14
SentiStrength+ TwitterSentiment	SNAP twitter dataset	69.7
SentiStrength	MySpace dataset	62.3
TwitterSentiment	SNAP twitter dataset	57.2

V. CONCLUSION

We have studied various approaches for sentiment analysis using machine learning techniques like Naive Bayes, SVM etc. The researches have done the summarization of events, real time event detection as well as sentence based sentiment classification accurately and efficiently. Naive Bayes classifier is insensitive to unbalanced data which give more accurate results.

VI. FUTURE SCOPE

We can use the PESTEL approach to perform the sentiment analysis on various domains separately i.e. to cluster all the tweets related to specific domain rather than mixed tweets. Here we can use the Support Vector Machine algorithm for sentiment classification which gives more efficient and accurate results as compared to sentiment analysis tools.

REFERENCES

[1] T. Sakaki, M. Okazaki, and Y. Matsuo, *Earthquake shakes twitter users: Real-time event detection by social sensors*, in Proc. 19thInt. Conf. WWW, Raleigh, NC, USA, 2010.

[2] D. Chakrabarti and K. Punera, *Event summarization using tweets*, in Proc. 5th Int. AAAI Conf. Weblogs Social Media, Barcelona, Spain, 2011.

[3] Y. Hu, A. John, F. Wang, and D. D. Seligmann, *Et-lda: Joint topic modeling for aligning events and their twitter feedback*, in Proc.26th AAAI Conf. Artif. Intell. Vancouver, BC, Canada, 2012.

[4] Asmaa Mountassir , Houda Benbrahim, Ilham Berrada, *An empirical study to address the problem of unbalanced data sets in sentiment classification*,IEEE International Conference on Systems, Man, and Cybernetics October 14-17, 2012, COEX, Seoul, Korea.

[5] Shulong Tan, Yang Li, Huan Sun, Ziyu Guan, Xifeng Yan, *Interpreting the Public Sentiment Variations on Twitter*, IEEE Transactions on Knowledge and Data Engineering, VOL. 26, NO.5, MAY 2014.

[6] Alexandre Trilla, Francesc Alias, *Sentence-Based Sentiment Analysis for Expressive Text-to-Speech*, IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 21, NO. 2, FEBRUARY 2013.

[7] Rui Xia, Feng X, Chengqing Zong, Qianmu Li, Yong Qi, Tao Li, *Dual Sentiment Analysis: Considering Two Sides of One Review*, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 8, AUGUST 2015.

[8] Rasheed M. Elawady, Sherif Barakat, Nora M.Elrashidy,"*Different Feature Selection for Sentiment Classification*, "International Journal of Information Science and Intelligent System, 3(1): 137-150, 2014.

[9] Duyu Tang, Bing Qin, Furu Wei, Li Dong, Ting Liu, Ming Zhou, "*A Joint Segmentation and Classification Framework for Sentence Level Sentiment Classification*,"IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 23, NO. 11, NOVEMBER 2015.

[10] X.Wang, F.Wei, X. Liu, M. Zhou, M. Zhang,"*Topic sentiment analysis in twitter: A graph-based hashtag sentiment classification approach*," in Proc.20th ACM CIKM, Glasgow, Scotland, 2011.

[11] Brendan O'Connor, Ramnath Balasubramanyan,"*From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series*," Proceedings of the International AAAI Conference on Weblogs and Social Media, Washington, DC, May 2010.

[12] M. Hu, B. Liu,"*Mining and summarizing customer reviews*,"Proc. 10th ACM SIGKDD, Washington, DC, USA, (2004).

.