RESEARCH ARTICLE OPEN ACCESS

Sentiment Analysis on Twitter Data Using Support Vector Machine

Bholane Savita D., Prof.Deipali Gore
Department of Computer Science and Engineering
Savitribai Phule, Pune University
Pune - India

ABSTRACT

Technology leads to advancements in research works. In today's technical age, e-commerce had been evolved so rapidly that millions or trillions of users are buying goods and services online over internet. One of the most important parts of e-commerce is sentiment analysis. One of the most visited social networking sites by millions of users is twitter. Here they put their opinion about various topics like politics, brands, products, and celebrities' etc. Variety of 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 contributed in finding reasons behind the sentiment variations and hence not useful in decision making. It has a big effect on the business intelligence field. In case when the seller wants to know why people are not buying his product, it becomes very difficult to survey customers who don't buy it. Hence seller uses sentiment analysis to search the web for opinions and reviews of this and competing products by using Blogs, Amazon and twitter like microblogging sites. Previously research was carried out to model and track public sentiments. But with the advancement in research, today we can use it for mining and summarizing products reviews, to solve the polarity shift problem by performing dual sentiment analysis.

Only modeling and tracking the public sentiments is not useful for decision making. Hence we use latent dirichlet based approach for sentiment variation tracking. It interprets exact reasons behind sentiment variations and ranks the reasons according to number of tweets.

Keywords:- Opinion mining, Latent Dirichlet Allocation, Public sentiment, emerging topic mining, FB-LDA, RCB-LDA, foreground topics.

I. INTRODUCTION

Sentiment Analysis is the process of finding the opinion of user about some topic or the text in consideration. It is also known as opinion mining. In other words, it determines whether a piece of writing is positive, negative or neutral. Now-a-days, people use microblogging sites to express their opinion about something. There are many popular microblogging sites like Facebook, Amazon etc. It has been useful in various domains like political, business and educational domain. Companies have been receiving polls about the products they manufacture. Previous research was to classify the sentiments into two classes i.e. positive and negative. But it was not useful for decision making. Here decision making refers to the solution for improving the positive opinion of the user regarding the domain in consideration. Hence need was to find the possible reasons behind sentiment variations to make decisions properly. E.g. if negative opinion towards Devendra Fadnavis increases significantly, Indian Administration maybe eager to know why people have changed their opinion from positive to negative so as to reverse the situation. There are many such examples in various domains like bollywood, political, health-care and business domain.

It seems very difficult to find the exact reasons behind sentiment variations as number of tweets are more than thousands for the target event. In this work, we find the sentiment polarity as well as reasons behind sentiment change for the target in consideration using two LDA models. We consider two tweet sets i.e. foreground tweets and background tweets (noisy data). To interpret the public sentiment variations, these background tweets should be removed. Tweets collection in the variation period consists of the main reasons as well as the neutral tweets i.e. background tweets which had been discussed from a long time. These are the tweets which are not contributing to the reasons for the change in public sentiment variations. Even the tweets are sometimes complicated and difficult to understand. The models are Foreground and Background LDA (FB-LDA) and Reason Candidate and Background LDA (RCB-LDA).

FB-LDA filters out the noisy data and distills the foreground topics from the tweets collection. To find out the foreground tweets, this model considers the noisy tweets before the variation period so as to filter the background tweets mixed with the foreground tweet collection. In this way the noisy data not contributing to the variations in tweets is removed using FB-LDA model. We get the optimized tweets related to the target. Now to find out the more relevant tweets for sentiment variations from tweets collection of FB-LDA, we use RCB-LDA model. It extracts the representative tweets as the reason candidates and then associates the remaining tweets with one reason candidate and ranks them by the number of tweets. In this way, using these topic models tracking and interpreting the public sentiment variations on

twitter data is done effectively for decision making applications.

a) Different Classes of Sentiment Analysis

There arethree classes of sentiments .i.e. positive, negative and neutral sentiments.

i. **Positive Sentiments**: This refers to positive attitude of the speaker about the text. Emotions with positive sentiments reflect happiness, joy, smile etc. In case of political reviews, if the positive reviews/sentiments about the politician are more, it means people are happy with his work.

ii. **Negative Sentiments**: This refers to negative attitude of the speaker about the text. Emotions with negative sentiments reflect sadness, jealousy, hate etc. In case of political reviews, if the negative reviews/sentiments about the politician are more, it means people are not happy with his work.

iii. **Neutral Sentiments**: Here no emotions are reflected about the text. It is neither preferred nor neglected. Although this class doesn't imply anything, it is very important for better distinction between positive and negative classes.

b) ***Levels of Sentiment classification:***

There are three sentiment classification levels. i.e. phrase level, document level and aspect level sentiment classification.

i. **Phrase Level Classification**: Phrase refers to combination of two or more than two words. Phrase is taken is consideration and sentiments are classified accordingly. But it sometimes gives inaccurate results due to addition of negation word. It first determines whether the phrase is neutral or polar. If it is polar then classified into positive and negative classes.

ii. **Document Level Classification**: it takes into consideration a single topic and classifies sentiment as positive, negative, or neutral. Sometimes it is not useful when there comparison between two products which have similar features.

iii. **Aspect Level Classification**: It determines whether the expressed opinion about the aspect or feature is positive, negative or neutral. Here aspect refers to the component of the entity. This classification level yields very fine grained opinion information which can be useful for various domains in sentiment analysis. The overall opinion is associated with the feature of the entity. E.g. entity maybe apple iphone and aspect maybe its battery, camera, screen etc.

II. LITERATURE SURVEY

Sentiment analysis is the most important research area in various fields. Different domains for sentiment analysis were educational, political, environmental, and social etc. Pang et al. has first carried out the basic work of sentiment classification in different areas. They have used star ratings as polarity signals in the training dataset.

A. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series

Brendan O'Connor proposed the model [1] in which measures of public opinion derived from polls with sentiment are measured from analysis of twitter data. We explicitly link measurement of textual sentiment in micro blogosphere messages through time, comparing to coeval polling data.

Data set:

They have considered 597 positive examples for classification of twitter data. It reports earthquake occurrence as a training set.

Advantages:

1. Summary statistics derived from different simple text analysis techniques are demonstrated to correlate with polling data on consumer confidence and political opinion.
2. It can also predict future movements in the polls.

Disadvantages:

1. It cannot either use tested tuning method or select the parameters of kernel based algorithms.
2. These works are not useful for removing noisy text data.
3. As explicit queries are not present in task, ranking methods cannot solve the reason mining task.

B. Different Feature Selection for Sentiment Classification

Sentiment analysis is a widely used research area for academic as well as business purposes but it faces the problems of existence noisy data. In this work, Rasheed E. et al. [2] have performed the task of feature selection using various methods. They have used three different techniques i.e.mRMR, IG and RST based feature selection. Hybrid feature selection based on IG and RST is the best technique than the IG for feature selection. The feature selection is done using unigrams. The results are tested on four datasets using two classifiers i.e. Naive Bayes and Support Vector Machine. SVM has best results as compared to NB classifier.

Data set:

The experiments are performed on two data sets i.e. movie review dataset and product review dataset. Both the datasets are consisting of 1000 positive and 1000 negative reviews for the classification.

Advantages:

1. Irrelevant and noisy features are eliminated properly from the feature vector for efficient working of Machine Learning algorithms.
2. SVM outperforms in sentiment classification than Naive Bayes classifier.

Disadvantages:

1. For the limited size of the training data, the system does not work properly.
2. For a very big dataset, it is a tedious task to select the appropriate features for machine learning and sentiment classification.

C. *Interpreting the Public Sentiment Variations on Twitter*

Twitter sentiment analysis is an important research area for political as well as business fields for decision making like for the administration to check whether people are happy with the work development of the president of the nation. In this work, Shulong Tan et al.[3] have proposed two models based on latent dirichlet allocation to interpret the sentiment variations on twitter i.e.FB-LDA to distill out the foreground topics and RCB-LDA to find out the reasons why public sentiments have been changed for the target.

Data Set:

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:

1. Uses the sentiment analysis tools TwitterSentiment and SentiStrength whose accuracy is less as compared to other sentiment analysis techniques.
2. It has been proved in the recent research [4] that SVM outperforms than the sentiment analysis tools used and other techniques in terms of accuracy and efficiency. Hence we prefer SVM for sentiment classification.

D. *Classification of Sentimental Reviews Using Machine Learning Techniques*

Abinash Tripathy et al. [4] have performed sentiment analysis on movie review dataset. They have compared results of different classification algorithms.

Data Sets:

Labeled polarity movie dataset.

Advantages:

1. SVM classifier outperforms every other classifier in predicting the sentiment of a review or tweet.

Disadvantages:

1. Only two different classifiers have been implemented and compared.

III. ARCHITECTURAL DIAGRAM

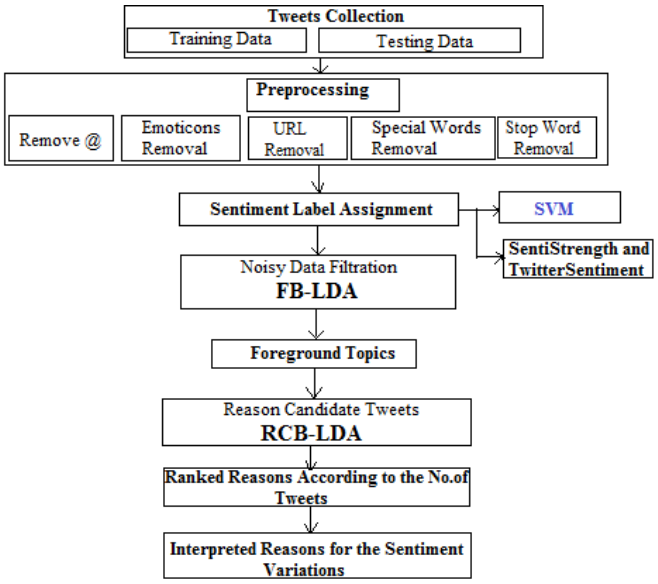

Fig.1.Sentiment Analysis on Twitter Data Using SVM

IV. ALGORITHMIC APPROACH

1. Tweets collection related to target in consideration.
2. Preprocessing of the data.
3. Sentiment label assignment using SentiStrength and Twitter Sentiment(later we have used SVM algorithm to improve the efficiency of results)
4. Tracking of sentiments about the target.
5. FB-LDA model to filter out background topics and foreground topic distillation.
6. RCB-LDA to extract representative tweets for foreground topics as reason candidates and associate each remaining tweet in variation period with one reason candidate and rank the reason candidates according to number of tweets associated with them.
7. Interpret the public sentiment variations about the target for decision making applications

a) SentiStrength and Twitter Sentiment tools:

Sentiment labeling is done using two tools, SentiStrength and Twitter Sentiment. Best results can be achieved using combination of the two. SentiStrength is based on LIWC library which assigns sentiment score to each word in the text. It computes sum of maximum positive and maximum negative score and assigns the positive, negative or neutral polarity as per the sign of the final score.

Twitter Sentiment is based on Maximum Entropy classifier. It chooses automatically collected noisy labels to train the classifier and based on the output, assigns the sentiment label.

b) Support Vector Machine(SVM):

SVM is generally used for text categorization. It can achieve good performance in high-dimensional feature space. An SVM algorithm represents the examples as points in space, mapped so that the examples of the different categories are separated by a clear margin as wide as possible. It gives best results than Naive Bayes algorithm and sentiment classification tools. 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.

V. MATHEMATICAL MODEL

Let S be the system that describes the tweet extraction, Preprocessing, Sentiment labeling, Sentiment Variation Tracking and Reason candidate for Sentiments.

- $S= \{T_w, P_t, S_l, V_t, R_s\}$
- T_w =Tweets extracted from Twitter.
- P_t =Preprocessing of Tweets (Slang word translation,Non-English word removal,PoS tagging,URL and Stop word removal).
- S_l=Sentiment Labeling using SentiStrength and TwitterSentiment sentiment analysis tools (SVM to give more efficient and accurate results).
- $S_l=\{P_v, N_v, N_e\}$
- $P_v=\{P_1,P_2,\ldots,P_n\}$= Positive Class
- $N_v=\{ N_1,N_2,\ldots,N_n \}$= Negative Class
- $N_e=\{ Ne_1,Ne_2,\ldots,Ne_n \}$= Neutral Class
- V_t =Sentiment Variation Tracking
- R_s=Reason Candidate for Sentiment Variations using RCB-LDA model

VI. EXPERIMENTS

A. Dataset Used

Dataset (D) = twitter data obtained from Sentiment 140(Twitter Sentiment) Platform.

D1 = Training dataset [1600000 tweets]

D2 = testing dataset [4000 tweets]

The dataset is in CSV file format where data is saved in table structured format. CSV file is used with spreadsheet program Microsoft Excel.

B. Performance Parameters for Evaluation

		Correct labels	
		Positive	Negative
Classified labels	Positive	True Positive(TP)	False Positive(FP)
	Negative	False Negative(FN)	True Negative(TN)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

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

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

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

Table 1. Performance Parameters

C. Results

Table.2.Shows original tweets with labeled sentiments. The tracked sentiments are in positive, negative and neutral classes. For the task of sentiment analysis, we have used Support Vector Machine algorithm that estimates the strength of positive and negative sentiment in short texts, even for informal language sentiments are classified on the basis of adjectives.

Text	Polarity
@Kelsie_love really, my boyfriend isn't going to get any siblings? whatevas, you're totally going to have another one, bitch. FRODO! baha	Negative
@littledotty nope not my job, not heard anything from that nor my assignment results im expecting, just about life really its poo	Negative
@tessdejong too bad its not MY bday party! hehe	Negative
Is working and then wedding at 2 yippy skippy	Neutral
Oh. Wow. dont look at people or scream conor oberst	Neutral
@xia_hime i'm going to aya! got a room onsite though so cant hotel share	Neutral
Finished garden for today. Hey fever broke out finally	Negative
@sebby_peek I kind of have to, for myself you don't like me though sweet dreams love xoxoxoxoxox	Positive
@drewyanscott nope but i want one to	Neutral
@UNO_OUT what about me jay I want a keychain	Neutral
@thepistol good luck with ur concert in san diego! wish i could be there but i live all the way in buffalo i know u'll do great tho! duh.	Positive

Table 2. Sentiment labeling using SVM

Fig. 2 shows the graph of sentiment variations. The X-axis consists of tweets and the Y-axis has the variations in sentiments. The varied sentiments can be positive, negative and neutral. Here we have entered the topic whose sentiment analysis to be done. We get the count of sentiment variations and the graph.

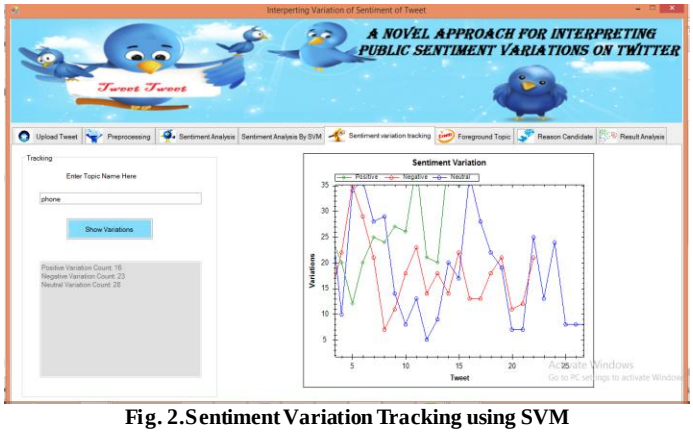

Fig. 2.Sentiment Variation Tracking using SVM

Table 3. Shows the table of foreground topics contributing in sentiment variation tracking. These are the most important words in tweets related to target.

Phone	cannot, receive, messages, anymore, ankle, answer, phones, reason, printer, wouldn't, accidently, plugged, earphones, battery, hangout

Table. 3. Foreground Topics

Table 4. Shows the table of reason candidates for the sentiment variations to be caused. It is done with the help of RCB-LDA model. RCB-LDA first extracts representative tweets for the foreground topics as reason candidates. Then it associates each tweet in the variation period with one reason candidate and ranks the reason candidates by the number of tweets associated with them. Here we have taken target as "phone" and we get the reason candidates about phone in descending order.

Reasons	Count
answer the phone, you bastard	238
What's worse than dropping your MacBook Pro? Dropping it onto your ...	226
just got home from haorizons golf thing now trying to get twitter on my ...	170
Hurt my ankle. Have to answer phones	154
Phone dropped in pond On approach. No more twittering on this trip	148
Can I drop my phone even more? No white screen yet.	144
@vrikis ages about about a year haha, i got totally conned over the ph...	141
someone stole my iphone lastnight! im devastated i cant beleive peop...	129
my phone is gonna die... no fair	129
today sucked im going to bed now lexi im going to call you from my da...	129
I tried setting up twitter on my phone but failed	118
Trying to find the receipt for the cellphone that's still not working	101
Death cab was sooo goood. my boyfriends not answering his phone. g...	93
My phone is still broken. I'm camping with work again so it won't really ...	92
I dropped my phone in a toilet. It's slowly RIPing. See ya later worlds.	92
@Katie_McElroy hheeeellllppp! i lost my phone and we need some pl...	84
trying to telll my friends to get a twitter!! my cell phone is dead!	82

Table.4. Reason candidate ranking for Sentiment Variation Detection of topic phone

Fig.3. shows precision-recall graph for sentiment analysis. Using these performance parameters, we get the accuracy of SVM is 97.54%.

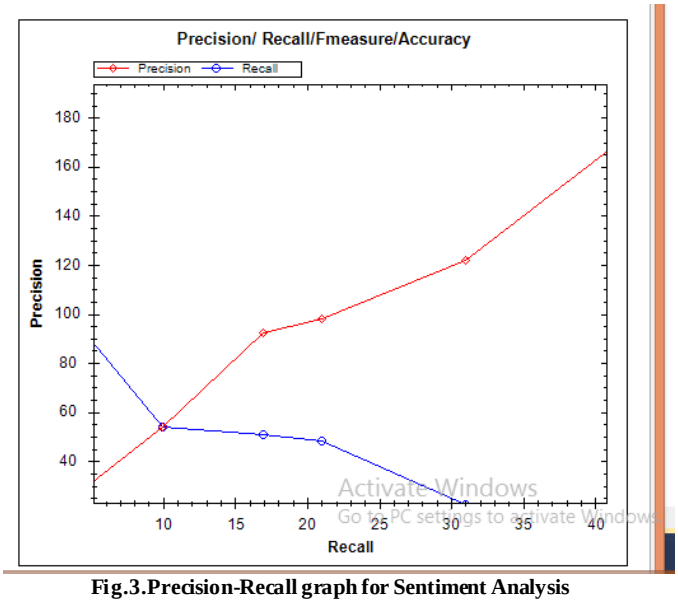

Fig.3.Precision-Recall graph for Sentiment Analysis

VII. COMPARATIVE RESULTS

Model/Algorithm/ Tools	Dataset	Accuracy (%)
SVM	Amazon product review data and ChnSentiCorp dataset	97.54
SentiStrength+ Twitter Sentiment	SNAP twitter dataset	74.30
SentiStrength	MySpace dataset	62.3
Twitter Sentiment	SNAP twitter dataset	57.2

Table.5.Comparative Results for Sentiment Classification Techniques

VIII. CONCLUSIONS

We have performed sentiment analysis on twitter data using two tools i.e. SentiStrength and Twitter Sentiment and support vector machine (SVM). Accuracy of interpretation of reasons behind sentiment variations is increased by 23.24% using SVM than that of the two tools. Hence we conclude that SVM is the best classifier for sentiment analysis.

Here we have interpreted most accurate reasons behind the change in sentiments of users using two LDA models, FB-LDA and RCB-LDA and SVM.

ACKNOWLEDGMENT

It is my privilege to express my sincerest regards to my M.E. coordinator, Prof. (Ms) D.V.Gore for her valuable inputs, able guidance, encouragement, whole-hearted cooperation and constructive criticism throughout the duration of the project work. As my guide, Professor D.V.Gore worked closely with me during the proposal completion.

I deeply express my sincere thanks to our Head of Department Prof.Dr.(Mrs)S.A.Itkar for encouraging and allowing me to present the project on the topic **Sentiment Analysis on Twitter Data Using Support Vector Machine** at department premises. I take this opportunity to thank all our lecturers who have directly or indirectly helped in my project work.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] Abinash Tripathy, Ankit Agrawal, Santanu Kumar, "Classification of Sentimental Reviews Using Machine Learning Techniques," *3rd International Conference on Recent Trends in Computing, 2015*.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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