Opinion Mining and Sentiment Analysis on Twitter

Aarti Gangawane
Dept. of Computer Science and Engineering
V.V.P. Institute of Engineering and Technology,
Soregaon, Solapur, Maharashtra, India

H. B. Torvi
Assistant Professor
Dept. of Computer Science and Engineering
V.V.P. Institute of Engineering and Technology,
Soregaon, Solapur, Maharashtra, India

ABSTRACT
In today's world of Internet, people became more eager to express and share their opinions on web regarding day-to-day activities and global issues as well. In this world of Internet various communication forms have emerged, by using these communication forms people can express their ideas, opinions about things which are going around them. Twitter is one those communication forms which are getting popularity now days among the people for interacting and expressing opinion in online word. Tweets on twitter provides an idea about peoples reaction toward particular event, product etc. This paper describes how sentiment analysis is done using tweets with text and emoticons. Tweets are collected for five different topics for implementing topic level search instead of the keyword based by using LDA algorithm. For sentiment analysis Bayesian logistic regression algorithm is used. By using twitter API tweets will be collected and after preprocessing the collected tweets classifier is applied to find the peoples sentiment towards given topic are positive, negative, strongly positive, strongly negative, or neutral.

General Terms
Twitter, tweets, Twitter API, Machine learning algorithm.

Keywords
Opinion mining, sentiment analysis, Bayesian logistic regression algorithm, LDA algorithm

1. INTRODUCTION
Sentiment analysis and opinion mining is the field where one can examine the people's reactions, emotions, opinions, sentiments, attitudes from the written language. It is the very interesting research area in natural language processing and also widely studied in text, data and web mining. The importance of opinion mining and sentiment analysis coincides with the evolution of social media such as blogs, reviews, forum discussions, microblogs, facebook, twitter etc. Sentiment analysis is being applied in every social and business domain because opinions are the key influencers of human behaviors as these are central to all human activities [21].

Now a day people are using microblogging sites to share their opinions and express their ideas about any topic, event or product [1]. It helps us to make analysis of that product, event etc. One of that microblogging a site twitters is getting more popularity for sharing the opinions. Tweets are the short messages used on twitter about 147 characters long to express ideas, opinions and events captured in the moment. This paper examine the commonly used machine learning method for text categorization which is Bayesian logistic regression [11] to find the positive, negative, strongly positive, strongly negative or neutral sentiment on tweets. And also will use LDA algorithm [15] for topic level search of tweets instead of keyword based search. We have collected tweets of five different topics for sentiment analysis. Tweets from twitter API will be collected for five different topics, which may contain either text or emoticons that are converted into text symbols accordingly and we will use these tweets as input to our system. An external lexicon such as Senti Word Net is used to support sentiment classification and opinion mining [19].

This paper is having sub sections are as follows. Which start first with an Introduction, in Section 2 discusses related work in Opinion Mining and Sentiment Analysis on Twitter. Proposed system and its system architecture in Section 3. Section 4 defines algorithm. The experimental results in section 5. Finally Section 6 represents conclusions and feature works.

2. RELATED WORK
Peiman Barnaghi, John Barslin and Parsa Ghaffari [2] gave a broad overview of Naive Bayes and BLR machine learning methods for sentiment analysis. They used FIFA world cup 2014 as a case study to find the sentiment of people towards that event are positive, negative, or neutral. Opinions have two polarities either positive or negative and if there is a lack of opinion it is neutral. This kind of labeling helps us to summarize the contents of document. So we need wide range of features for opinion and polarity detection.

Feldman et al [1] has given three levels of sentiment analysis: document-level sentiment analysis, sentence-level sentiment analysis and aspect based sentiment analysis. Document-level sentiment analysis focuses on single entity or event, the opinion here are classified into simple classes: positive or negative (probably neutral).Sentence-level analysis have more refined view of different opinions expressed in the document about the entities. Aspect based analysis refers to recognition of all sentiment expressions within a given document and the aspect to which the opinion refers.

Many advanced methods and algorithms have been developed for text categorization during the last three decades [6].The bag-of-words method is a standard approach and the most popular model for text categorization [18] as the concept is easy to understand and also helps improve performance. The bag-of-words method uses a vector of words in Euclidean space for representing the document where each word is independent from others and used as a feature for training a sentiment model [18]. R. Basili, A. Moschitti, and M. T. Pazienza [6] in 2000 gave new way of data representation which was specifically designed for text representation. They used kNN classifier for document categorization. This representation solves the high dimensionality problem for large number of dataset.

3. PROPOSED SYSTEM
Figure 1 shows the system architecture of proposed system for sentiment analysis. Figure shows different modules and steps of tweets preprocessing, feature extraction and feature filtering, to get a trained classifier for sentiment polarity

classification. The system architecture of proposed system shows different modules such as,

Tweets collection module: Collected tweets from twitter API [10] contains useless information because of this the workflow is designed in order to clean using preprocessing such as tokenization, stopwords removal, upper case conversion, stemming and lemmatization, negation handling and also converting the contents of tweet massage such as username,URLs to general tags and hashtags to mark the topic.

Tweets preprocessing module: Tweets preprocessing module is responsible for breaking the sentence into words. Tokenization, stemming [17], twitter symbols and hashtags etc. are performed in this module.

Feature extraction and feature filtering: In feature extraction and filtering module LDA algorithm is used to get more refined input for sentiment analysis. Useful words and features are extracted to calculate the sentiment result by applying TF-IDF [20].

Classifier: Finally the classifier module is applied to calculate accuracy of sentiment analysis which gives the sentiment result as positive, negative, or neutral.

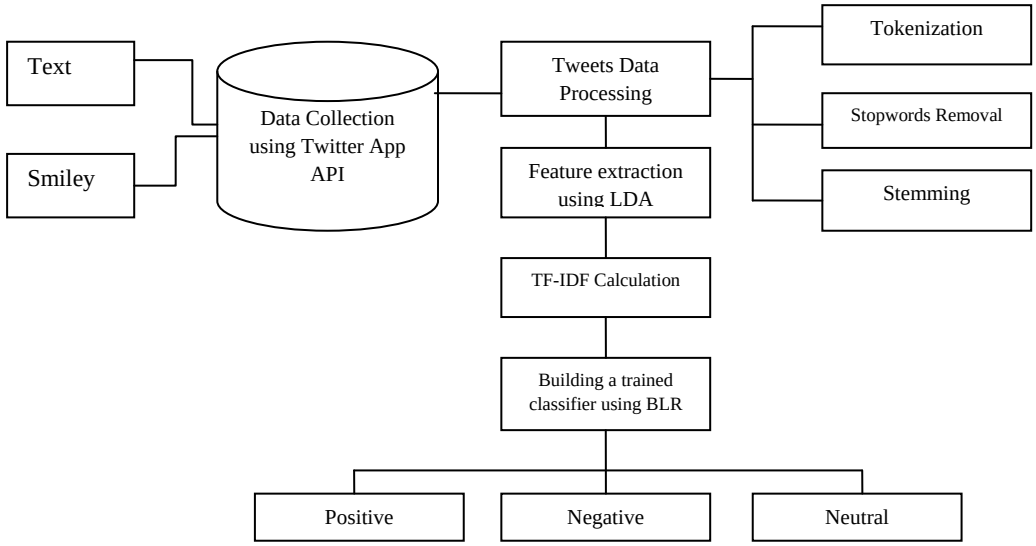

Fig 1: Opinion Mining And Sentiment Analysis System using LDA and BLR

4. ALGORITHM

4.1 LDA Algorithm

Proposed system uses topic level search instead of keyword based search, for this LDA algorithm (Latent Dirichlet Allocation) is used. The main purpose of topic level modeling is to recognize the patterns of words that are used to connect the same documents that shares same patterns [14].

LDA is the probabilistic model amongst various topic generative models which are used to analyze large collections of text corpora, where each document is the collection of words and each word is generated from certain topic which is drawn from topic distribution. The topic is latent distribution of document and can be viewed as a feature for prediction purpose i.e. Sentiment analysis. The generative process of LDA is as follows [15], for each of N_j words in document j

1) Choose a topic $Z_{i,j} \sim Mult(\theta_j)$

2) Choose a word $X_{i,j} \sim Mult(\varphi Z_{i,j})$

The parameters of multinomial for topics in the document θ_j and words in a topic φ_k have Dirichlet priors [13].

4.2 Bayesian Logistic Algorithm

Bayesian logistic regression model is used as classifier in text categorization, by selecting features simultaneously.BLR selects features and provides optimization for performing text categorization [16]. BLR avoids the problem of over fitting by using Laplace prior formula and produces sparse predictive model [11]. The estimation of BLR is as follows.

$$P(c|f) = \frac{1}{z(f)} \exp\left(\sum_i \lambda_{i,c} F_{i,c}(f,c)\right)$$

Where z (f) is a normalization function, λ is a vector of weight parameters for the feature set [20], and $F_{i,c}$ is a binary function that takes the input the feature and a class label [1].

$$F_{i,c}(f,c') = \begin{cases} 1, & n(f)>0 \text{ and } c'=c \\ 0, & \text{otherwise} \end{cases}$$

This binary function is getting called when any feature is encountered such as unigram feature or bigram feature and the sentiment is predicted anyway.

5. EXPERIMENTAL RESULTS

TF-IDF and Sentiments Result: Sentiment classifier gives the accuracy of sentiments towards particular topic. Proposed system uses LDA for topic level search which makes the BLR to select more accurate features for sentiment polarity detection. Below is the tabular representation of TF-IDF, positive score, negative score, strongly positive score, strongly negative score and neutral score.

Table 1: TF-IDF and Sentiments Result

Topic Name	TF-IDF	Positive Score	Strong Positive Score	Negative Score	Strong Negative Score	Neutral Score
GST in India	0.2578	0.6091	1.6160	-0.4198	-0.4229	0.1451
Demonetization	0.1546	0.3453	0.4138	-0.4721	-1.0503	0.011
FIFA World cup 2014	0.1164	0	0	-0.5847	-0.2936	0.00015
Avengers	0.4707	1.7610	5.4445	-1.8038	-1.7095	0.3202
Plastic Ban	1.1048	0.0361	0.3070	0	0	0.1229

The below graph indicates sentiment analysis of five different topics which shows result in five different categories that is positive, negative, strongly positive, strongly negative or neutral.

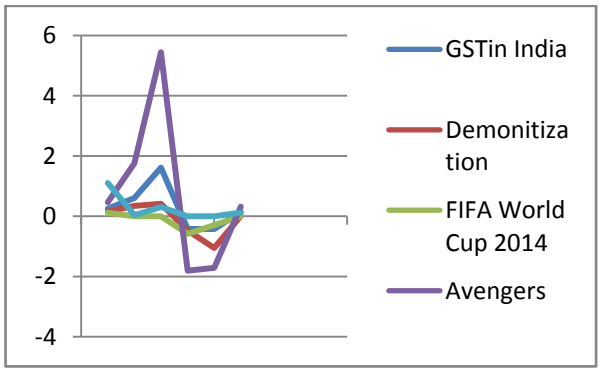

Fig 2: TF-IDF and Sentiments Result

Sentiment Result Accuracy: From the searched topics retrieved tweets are used to calculate Sentiment result accuracy which is shown below in table.

Table 2: Sentiment Result Accuracy

Topic Name	Total Tweets	Positive Tweets	Negative Tweets	Neutral Tweets	Accuracy in %
GST in India	125	34	22	69	71.79%
Demone-tization	82	18	23	41	64.06%
FIFA World Cup 2014	126	32	40	54	64.29%
Avengers	328	57	61	210	65.92%
Plastic Ban	12	1	0	11	98%

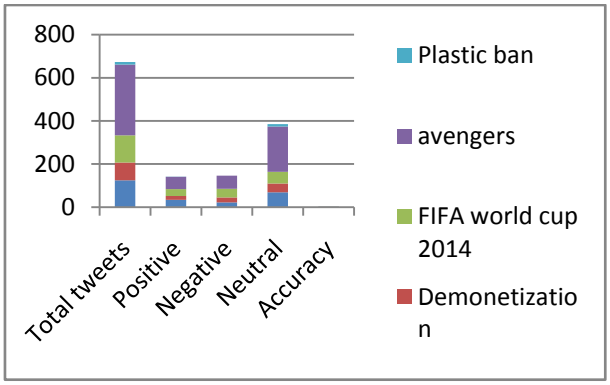

Fig 3: Sentiment Result Accuracy

The above graph explains Sentiment result accuracy in percentage. For searched topics total number of tweets collected, Positive tweets, negative tweets, neutral tweets are calculated.

Evaluation Matrix: The performance of implemented algorithm for topic selection (LDA) is evaluated on below measures.

Table1 3: Performance Evaluation and BLR Accuracy

Topic Name	Algorithm Accuracy in %	Precision	Recall	Fmeasure
GST in India	82.00	0.87879	0.56862	0.69047
Demonetization	81.00	0.761904	0.41025	0.53333
FIFA World Cup 2014	77.00	0.25	0.58823	0.09523
Avengers	72.00	0.7875	0.50806	0.61764
Plastic Ban	76.00	1	1	1

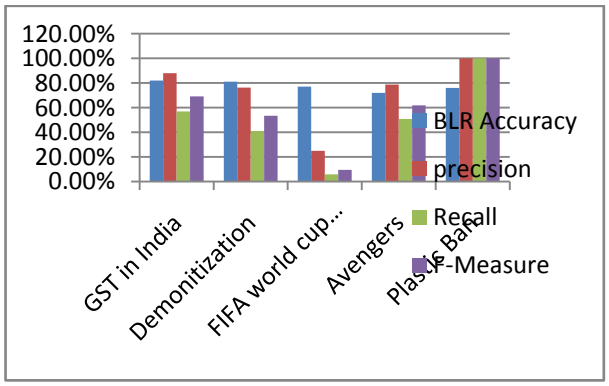

Fig 4 : Performance Evaluation and BLR Accuaracy

6. CONCLUSION

Twitter contains lots of noisy data sets; these data sets are processed to locate interesting trends. Opinion mining and sentiment analysis system is used to detect people's reactions, opinions, reviews towards particular topic, event, product etc. In proposed system Latent Dirichlet Allocation algorithm is used for topic level search of tweets which helps the Bayesian Logistic algorithm to give the more accuracy of sentiment classification. Tweets can contain either text or emoticons, both are handled well by LDA algorithm in topic level search of tweets. BLR gives the more accurate result of sentiment analysis as compare to other techniques as it has less time complexity than other techniques.

In future more work is needed to improve the performance. Sentiment analysis can be applied for new applications. Although the techniques and algorithms used for technology analysis are in progress, however, many problems in this field of study are not solved, such as challenge in handling other languages, handling complexity of documents or sentences etc. For this lot of future work could be dedicated.

7. REFERENCES

[1] P. Barnaghi, P. Ghaffari, and J. G. Breslin, "Text Analysis and Sentiment Polarity on FIFA World Cup 2014 Tweets," in Conference ACM SIGKDD, 2015.

[2] Peiman Barnaghi, and John G. Breslin," Opinion Mining and Sentiment Polarity on Twitter and Correlation between Events and Sentiment".2016 IEEE Second International Conference on Big Data Computing Service and Applications.

[3] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," in LREC, 2010, pp. 1320-1326.

[4] Melville, Wojciech Gryc, "Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification", KDD˝09, June 28–July 1, 2009, Paris, France.Copyright 2009 ACM 978-1-60558-495-9/09/06.

[5] Golbeck, Jennifer, et al. "Predicting personality from twitter." Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on. IEEE, 2011.

[6] R. Basili, A. Moschitti, and M. T. Pazienza. Language-sensitive text classification. In Proceedings of 6th International Conference "Recherché information Assistee par Orinateur", 331-343, 2000.

[7] Asiaee T, A., Tepper, M., Banerjee, A., and Sapiro, G. (2012). If you are happy and you know it... tweet. In Proceedings of the 21st ACM international conference on Information and knowledge management, pages 1602–1606. ACM.

[8] Gokulakrishnan, B., Priyanthan, P., Ragavan, T., Prasath, N., and Perera, A. (2012). Opinion mining and sentiment analysis on a twitter data stream. In Advances in ICT for Emerging Regions (ICTer), 2012 International Conference on, pages 182–188. IEEE.

[9] Figueiredo, M. A. T., and Jain, A. K. (2001), "Bayesian Learning of Sparse Classifiers," in Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1, pp. I-35–I-41.

[10] A. Bifet and E. Frank, "Sentiment knowledge discovery in twitter streaming data," in Discovery Science, 2010, pp. 1-15.

[11] A. Genkin, D. D. Lewis, and D. Madigan, "Large-scale Bayesian logistic regression for text categorization," Technometrics, vol. 49, pp. 291-304, 2007.

[12] R. Feldman, "Techniques and applications for sentiment analysis," Communications of the ACM, vol. 56, pp. 82-89, 2013

[13] Porteous, L., Newman, D., Ihler, A., Asuncion, A., Smyth, P., and Welling, M., ―Fast Collapsed Gibbs Sampling for Latent Dirichlet Allocation‖, ACM New York, NY, USA, 2008.

[14] Blei, D., Ng, A., and Jordan, M. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3:993–102.

[15] Blei, D.M., and Lafferty, J. D. ―Dynamic Topic Models‖, Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, 2006.

[16] David Zimbra, M. Ghiassi and Sean Lee, "Brand-Related Twitter Sentiment Analysis using Feature Engineering and the Dynamic Architecture for Artificial Neural Networks", IEEE 1530-1605, 2016.

[17] V. N. Vapnik and V. Vapnik, Statistical learning theory vol. 1: Wiley New York, 1998. [18] G. Salton and M. J. McGill, "Introduction to modern information retrieval," 1986.

[18] S.Dumais. J. Platt, D. Heckerman, and M. Sahami,"Inductive learning algorithms and representation for text categorization, "in Proceeding of the seventh international conference on Information and knowledge management, 1998, pp-148-155.

[19] S.Baccianella, A. Esuli, and F. Sebastian, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," in LREC, 2010, pp. 2200-2204.

[20] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up: sentiment classification using machine learning techniques," in Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, 2002, pp. 79-86.

[21] Akshi Kumar and Teeja Mary Sebastian, "Sentiment analysis on twitter", IJCSI, Vol. 9,Issue 4, No 3, July 2012 ISSN (Online): 1694-0814