ISSN No. 0976-5697

DOI: http://dx.doi.org/10.26483/ijarcs.v8i9.5212

Volume 8, No. 9,November-December 2017

International Journal of Advanced Research in Computer Science

SURVEY REPORT

Available Online at www.ijarcs.info

A COMPARATIVE STUDY OFSENTIMENTAL ANALYSIS IN VARIOUS TECHNIQUES

S.Indhu
Research Scholar,
Department of Computer Science
Sri Ramakrishna College of Arts and Science
for Women Coimbatore, India

Mrs.S.R.Lavanya
Assistant Professor,
Department of Computer Science
Sri Ramakrishna College of Arts and Science
for Women Coimbatore, India

***Abstract*:** Social media is one of the significant places to show user's feelings. In sentiment analysis the information is separated from the notion, assessment and feelings of individuals with respect to entities, events, and their characteristics. Sentiment analysis is also known as Opinion Topic Modeling (OTM) approach. OTM is used to examine and group the user created information likecomments, articles and so on. Thisinformation's are taken from the social networking sites like facebook, twitter and so on. Twitter has given a huge space for analysing movie reviews, consumer brands, election events, stock market exchange, and so on. This survey paper discusses several methods used for sentiment analysis.

***Keywords*:** Social media, Twitter, Topic modeling, LDA.

I. INTRODUCTION

The microblogging site liketwitter entices more users to post their sentiments and feelings on different topicscalled tweets [1]. Tweets are content-based posts with 140 characters long and it grants individuals to pursue and read each other's tweets [2]. The posting of sentimental words in online not only give emotional values and also have other values. But,it is difficult for individuals to get an overall impression from millions of tweets without automatic classification. As a result, there are several sentiment classification tasks indicating benefits in tweets.

Twitter allows different types of topics to discuss. Sentiment classifiers commit themselves to a particular area or topic. In particular, sentiment data from one subject frequently performs ineffectively on test data from another [3]. Because words and languageconstructs utilized for conveying feelings can be unique on various topics. For example, "Samsung is good" it could be confident in a product review but impropersentiment in a movie review. In social media, a Twitter client may have distinctive feelings on various subjects. In this way, the topic adjustment is required for sentiment classification of tweets on developing and random subjects. Past works expressly acquired a connection to interface a topic-dependent feature to a known or basic component. And the connections are made between item reviews by expecting that the similar sentimental words exist for each couple of topics. Though, it is not really appropriate to the topics in Twitter, particularly the random ones. It is value to saying that identifying and following topics from tweets are another research topic.

Sentiment analysis is helpful for aninvestigation into online correspondence because it enables researchers the ability to mechanically quantify feeling in online writings [4]. Generally, sentiments and opinions can be analysed at different levels of granularity. The task of analysing overall sentiments of texts is typically formulated as a classification problem, e.g., classifying a review document into positive or negative sentiment. Then, a variety of machine learning methods trained using different types of indicators (features) have been employed for overall sentiment analysis. However, analysing the overall sentiment expressed in a whole piece of text alone (e.g., review document), does not discover what specifically people like or dislike in the text. In reality, the fine-grained sentiments may very well tip the balance in purchase decisions.

Sentiment analysis remains a challenge in unsupervised models.Unsupervised techniques include determining a sentiment classifier without any categorized data [5]. While the sentiment is relatively easy to detect in supervised experiments, models lacking supervision are often unable to make the distinction properly. In some domains (for example,book surveys), topics firmly interfere with opinion and unsupervised models scarcely beat the chance baseline. Bayesian models have gotten significant consideration in natural language processing research. They empower the joining of earlier learning in arbitrary graphical models while parameter implication can be accomplished with simple sampling methods. This paper presents a survey of sentiment analysis techniques.

II. LITERATURE SURVEY

In [6] authors defined a probabilistic model Latent Dirichlet Allocation (LDA), is a model for grouping of separate data, for instance, text corpora. In LDA method each component of a group is revealed as a limited combination of afundamental set of themes. Every theme is showed as an immeasurable combination of afundamental set of subject possibilities. In the circumstance of text modeling, the topic possibilities contribute an obvious depiction of a text. The well-organized estimated implication methods depended on variation methods and EM algorithms for empirical Bayes parameter estimation is revealed in this paper. The outcomes of collaborative filtering, document modeling, and text classification are conveyed and likened to a combination of unigrams model and the LSI model.

In [7] an author explained about Twitter is a microblogging site in that people read and compose lots of short messages on an assortment of subjects each day. This experiment utilizes the circumstance of the German federal election to research whether Twitter is utilized as a discussion for political pondering and whether online messages on Twitter truly reflect offline political opinion. With LIWC text analysis software, they directed a gratified investigation of more than 100,000 messages covering a reference to either a political party or politician. The outcomes demonstrate that Twitter is really exploited widely for political thought. And found that the meagre amount of messages specifying a party imitates the election result.

In [8] authors examined the predictive analysis on web-based networking media time-series enables the stakeholders to use this immediate, accessible and vast reachable communication channel to respond and proact against the public opinion. Specifically, understanding and calculating the sentiment change of the public opinions will enable business and government organizations to respond against negative sentiment and design strategies, for example, dispelling gossips and post balanced messages to return the public opinion. The authors exposed a system of building statistical models from the social media dynamics to forecast collective sentiment dynamics. The Combined sentiments modify without investigating into a micro analysis of separate tweets or users and also their conforming low-level system structures.

In [9] authors developed a general answer for sentiment classification when there are no labels in the target domain but have some labeled information in an alternate domain, viewed as source domain. In this cross-domain sentiment classification setting, it ties the gap between the domains. It offers a spectral feature alignment (SFA) algorithm to align domain-specific words from various domains into unified groups, with the support of domain-independent words as a bridge. This way, the groups can be used to decrease the gap between domain-specific words of the two domains, which can be used to train sentiment classifiers in the target domain exactly. Compared with past methodologies, SFA can find a robust representation for cross-domain data by completely exploiting the connection between the domain-specific and domain independent words by simultaneously co-clustering them in a common latent space.

In [10] authors presented a new method for automatically classifying the sentiment of Twitter messages. The messages are divided into positive and negative groups with the help of the given query. This is helpful for customers who need to inquire about the opinions of items before buying, or organizations that need to monitor the public opinion of their products. There is no preceding research work on categorizing feeling of messages on Twitter. The conclusion of machine learning algorithms for categorizing the opinions of Twitter messages with distant supervision is presented in this paper. The training dataset consists of Twitter messages with emojis, which are taken as noisy labels. This sort of training information is plentifully accessible and can be originated through computerized implies. The machine learning algorithms (Naive Bayes, Maximum Entropy, and SVM) have precision over 80% when prepared with emoji data are appeared in this paper.

In [11] authors presented a method namely S^3VM to develop a support vector machine utilizing both the training and functioning sets. Here authors used S^3VM to solve the issue of transduction by Overall Risk Minimization (ORM). The transduction issue is to assess the significance of a grouping task at the specified points in the functioning set. This appears differently through the standard inductive learning issue of evaluating the classification function at all conceivable values and next utilizing the stable function to infer the classes of the functioning set information. This model reduces both the misclassification mistake and the practical capability depends on all the manageable information. This paper specified how the S^3VM model for 1-norm linear support vector machine can be transformed to a mixed-integer program and subsequently, that resolved precisely utilizing number programming. Consequences of S^3VM and the standard 1-norm support vector machine approach are likened to ten data sets. The outcome demonstrates that integrating functioning data expands simplification when inadequate training information is accessible.

TABLE 1: SUMMARY TABLE FOR COMPARISON OFTOPIC MODELING TECHNIQUES

Title	Algorithm	Merits	Demerits	Results	Performance
Latent Dirichlet Allocation [6]	Latent Dirichlet Allocation	The Probabilistic model that can be easily extended and embedded in other more complicated models.	Prone to over-fitting (have to be careful when training).	LDA is a flexible generative probabilistic model for collections of discrete data.	93 % proportion of data used for training.

Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment[7]	LIWC (Linguistic Inquiry and Word Count) text analysis software.	An analysis of the tweets' political sentiment demonstrates close correspondence to the parties.	The current lexicon not particularly customized to classify the short messages as tweets.	The resemblance of the profiles is suggestive of the parties' vicinity concerning political issues.	In this model,60% of all messages declared the name of the political party.
Predicting collective sentiment dynamics from time-series social media [8]	multiclass SVM (Support Vector Machine)	A strategy of building statistical models from the social media dynamics to predict collective sentiment dynamics.	Machine learning or other methods for automatic sentiment analysis is difficult.	SVM and logistic regression have a similar result and outperform the decision tree.	This model can accomplish over 85% precision
Cross-Domain Sentiment Classification via Spectral Feature Alignment [9]	Spectral Feature Alignment (SFA) algorithm	The clusters can be used to reduce the gap between domain-specific words of the two domains in the target domain accurately.	To address multi-category sentiment classification problems.	Outcomes of the sentiment classification tasks prove the efficiency of the structure.	86.75% Accuracy using SFA.
Twitter Sentiment Classification using Distant Supervision [10]	Naive Bayes, Maximum Entropy (MaxEnt), and Support Vector Machines (SVM).	The benefit of the reduction properties is to decrease the feature space.	Crumpling the separate words to equality classes does not support.	Results that unigram feature extractor is the easiest approach to recover features from a tweet.	95.7% Accuracy using Feature reduction process.
Semi-Supervised Support Vector Machines [11]	Semi-Supervised Support Vector Machine	SVM model that minimizes both the misclassification error and the functional capacity based on all the available data.	An issue with small training sets and large functioning sets is a type of semi-supervised clustering.	Outcomes show that incorporating functioning data enhances simplification when inadequate training information is accessible.	80.63% Accuracy.

III. CONCLUSION

Thissurvey paper discussed several methods used for sentiment analysis. Sentiment detection has a wide variety of applications in information systems, including classifying reviews, summarizing review and other real-time applications etc. Sentimental classifiers are subjected to particular domain or topic. The study shows that machine learning methods, such as SFA and naive Bayes have the highest accuracy and can be regarded as the baseline learning methods, while LDA method is very effective in some cases, which require little effort in the human-labeled document.The future work examines the behavior of the Naive Bayes model inan unsupervised setting where some document or feature labels are known.

REFERENCES

[1] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, “Sentiment analysis of Twitter data,” in Proc. Workshop Lang. Soc. Media, 2011, pp. 30-38.

[2] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, “Micro-blogging as online word of mouth branding,” in Proc. Extended Abstr. Human Factors Comput. Syst., 2009, pp.3859-3864.

[3] B. Liu, “Sentiment analysis and opinion mining,” Synthesis Lect. Human Lang. Technol., vol. 5, no. 1, pp. 1-167, 2012.

[4] M. Thelwall, K. Buckley, and G. Paltoglou, “Sentiment in Twitter events,” J. Am. Soc. Inform. Sci. Technol., vol. 62, no. 2, pp. 406-418, 2011.

[5] S. Li, C.-R. Huang, G. Zhou, and S. Y. M. Lee, “Employing personal/impersonal views in supervised and semi-supervised sentiment classification,” in Proc. 48th Annu. Meeting Assoc. Comput. Linguistics, 2010, pp. 414–423.

[6] D.M. Blei, A.Y. Ng, and M.I. Jordan, “Latent Dirichlet Allocation,” J. Machine Learning Research, vol. 3, pp.993-1022, 2003.

[7] A. Tumasjan, T.O.Sprenger, P.G. Sandner, and I. M. Welpe, “Predicting elections with Twitter: What 140 characters reveal about political sentiment, “ in Proc. 4th Int. AAAI Conf. Weblogs Soc. Media, 2010, vol. 10, pp.178-185.

[8] L. T. Nguyen, P. Wu, W. Chan, W. Peng, and Y. Zhang, “Predicting collective sentiment dynamics from time-series social media,” in Proc. 1st Int. Workshop Issues Sentiment Discovery Opinion Mining, 2012, p.6.

[9] S. J. Pan, X. Ni, J.-T. Sun, Q.Yang, and Z. Chen, “Cross-domain sentiment classification via spectral feature alignment,” in Proc. 19th Int. Conf. World Wide Web, 2010, pp. 751–760.

[10] A. Go, R. Bhayani, and L. Huang, “Twitter sentiment classification using distant supervision,” CS224N Project Report, Computer Science Department, Stanford, USA, pp. 1–12, 2009.

[11] K. Bennett and A. Demiriz, “Semi-supervised support vector machines,” in Proc. Adv. Neural Inform. Proc. Syst., 1999, pp. 368–374.