SECURED CLASSIFICATION OF SENTIMENTS ON TOPIC ADAPTIVE DYNAMIC TWEETS

Shrivatsa D Perur[1], Bhavya Balakrishnan[2]

[1]*Department of CSE, AMCEC, Bengaluru,*
perur35@gmail.com
[2]*Department of CSE, AMCEC, Bengaluru,*
bhavya.balakrishnan@gmail.com

Abstract

The work of classifying sentiments is adaptive to subject, a classifier prepared to perform on a topic will not have same effect on other. This poses a hindrance for the analysis of sentiments. There will be various topics in Twitter, which makes the task difficult for preparing a generalized classifier for all subjects. However, when comments on item is considered, data labelling is not provided in micro blogging site.furthermore, a rating component to obtain conclusion names. Here, we propose a semi-managed notion arrangement (SC) model, which begins with a classifier, based on basic components and blended named information from different subjects. It minimizes the pivot misfortune to adjust to unlabeled information and components including subject related notion words, creators' conclusions and opinion associations got from "@" notice of tweets, named as point versatile elements. Content and non-content components are extricated and normally split into two perspectives for co-preparing. Classified tweets are maintained securely.

Keywords*: Twitter; Classifier; Sentiments*

----------***----------

1. INTRODUCTION

Sentiment classification deals with sensitive data field. Twitter attracts huge number of individuals to put the reviews, assessments on different points. tweeting over topics not just gives enthusiastic depiction , in addition to that it gives a potential commercial, financial and sociological value [1][2][3][4]. It is very difficult for people to come to a conclusion based on the tweets because they are so massive that it becomes very difficult to analyse the sentiment over the product. Subjects talked in twitter are more different which cannot be predicted. The sentiment classifiers restrict themselves to a particular domain or a topic. A classifier trained to work on a particular topic will fail to work on another topic. The twitter user may have difference in opinion over a topic. For example, a person may give a positive feedback for a movie while another person may give negative feedback.

Thus, classification of sentiment of tweets on emerging and topics that cannot be predicted, topic adaption is needed.[5] Few works [6][7][8] in past have borrowed a link that connects feature that is topic dependent and a pre determined or common feature. However, such kind of links may not be applied over topics in twitter which are unpredictable. However micro blogging site needs labelled information and rating tool to be applied on it. [9] has made use of the emoticons as boisterous marks for feeling grouping. However this may not be used to label the neutral classes since noise may not be introduced through the emoticons only. [10][11][12][13] have made use of a semi supervised approaches to classify sentiments with a little measure of labelled information for other gateway than Microblog.

Emoticons, users etc were not taken to choose unlabeled data for training purpose. [5] explored that the correlation of sentiment are influenced by clients who are mutually connected with each other via social media. It has been well recognized that the content generated by the users that has rich sentiment is to be used for various applications and information systems. Though the sentiment analysis at tweet level gives useful information, the common tendency of sentiment towards a particular scenario is more appealing. For example, when a new cell phone has been launched, people want to know how others feel about the cell phone and this will help them to decide over things from massive response. Fans of celebrity would be keen on knowing what is going on in their favourite celebrity's life and how others respond to it. The analysis of comprehensive sentiment tendency is required towards a topic in such scenarios. To satisfy this demand, has made use of hashtag characteristic in twitter.

2. RELATED WORK

Sentiment classification on cross domain topics is a challenging task. An approach, Structural Correspondence Learning was proposed for domain adoption. This acts as a bridge for cross domain classification. Pan et al., [14] proposed an algorithm SFA to bridge gap between domain and domain independent words. Twitter data contains diverse topics from different domains and different topics which are unpredictable and labelling of data for each topic is needed. SUIT model[15] considers point perspectives and

sentiment holders for classification of sentiments through supervised learning. Twitter information set is not quite the same as different assets. Microblogs attracted huge studies on sentiment analysis as a social media[16][17][18][19] tasc]. For automatically classifying sentiments of noisy labels for data sets, a distant supervised learning approach is introduced. [4] showed twitter may be seen as a predictor for political opinion. We focus on sentiment classification problem of tweets. For supervised sentiment classification, lack of labelling is an issue. Visualizing themes[20][21][22] even in text mining domain[23].

3. MOTIVATIONS

Tweets that are publicly available with labels on diverse topics are considered. With necessary pre processes, the frequent adjectives, verbs, nouns and adverbs are selected for sentiment words as candidates. Different topics use different words for sentiments.

Table 1: Statistics

Topics	Positive	Neutral	Negative	Total
Apple	191	581	377	1149
Google	218	604	61	883
Microsoft	93	671	138	902
Twitter	68	647	78	793
Taco Bell	902	2099	596	3597
President Debate	1465	1019	729	3213

The tweets are divided into few different topics. The detailed data information is shown in table 1. To worsen the matters, same word may be having different sentiments for different topics. For example the word "unpredictable" has positive sentiment for few topics and negative for few topics. This hinders the sentiment classifiers to adapt themselves for different topics as such. Table 2 s sentiment words captured from tweets over variou The sentiment of a user over a context should be c over a context along with reflecting his opinion on it.

Direct parent

Table 2. Opinion words captured from tweets

Topics	Sentiment Words
Apple	Amazing, Better, Design, Genius, Great, Service
Google	Available, Cool, Unveil, Sharing, Infinite, Really
Microsoft	Available, Celebrity, Deal, Free, Learning, Review

If the sentiment of the user is evenly distributed over a topic, there is a chance of posting positive, negative and neutral tweets is equal. The variance of the sentiment of user is calculated using the formula,

$$Var(A) = M(A^2) - (M(A))^2$$

Suppose the tweets are evenly distributed, 1/3 is the probability of tweet being positive, negative or neutral is possible. In the above equation, M(.) is the mean and Var(.) is the variance of opinion of the tweets.

Finally, "@" is a commonly used convention of tweets. Such @ depicts the dependencies of tweets and the user to whom it is referenced.

Multiclass SVM

The model of SVM is built for binary classification. There are many ways to solve the multiclass with SVMs. One versus rest classifiers are the common method that has been incorporated to choose the class which classifies test data with greater margin. Here we build a one to one classifiers and choose a class that is chosen by most classifiers.

The model for SVM is as follows

$$\min_w \tfrac{1}{2} \Sigma_{i=1}^{N} w_i^T \; w_i + \tfrac{C}{n} \sum_{i=1}^{n} \max_{y \neq yi} 0, 1 - w_{yi}^T x_i + w_y^T x_i$$

The w in the equation is the matrix with w_i as coefficient vector referring to feature of class i $\in 1, \ldots\ldots, K$. $w^T_{yi} x_i$ is the confidence score of tweet t_i belonging to class y, C is constant coefficient. The equation shows that the auxiliary danger is used to improve and after effect of model is single vector machine rather than multiple one.

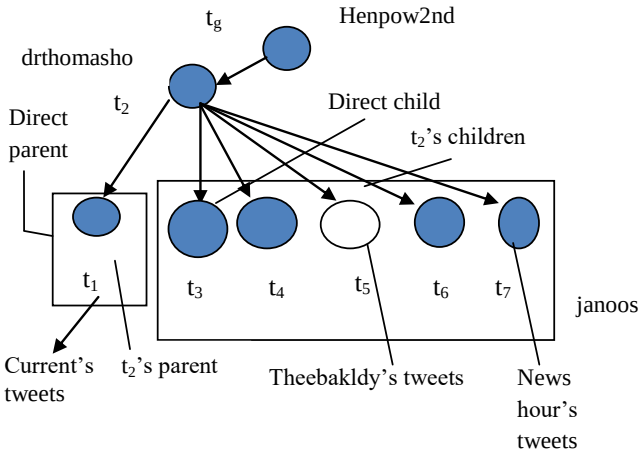

Fig 1: Example of @ network

	Post time	Author	@whom	Sentiment
t_1	2:01	Janoos	@current	Negative
t_2	2:13	drthamasho	@janoos	Negative
t_3	2:19	Janoos	-	Negative
t_4	2:29	Janoos	-	Negative
t_5	2:37	Janoos	@theebakldy's	Neutral
t_6	2:38	Janoos	-	Negative
t_7	2:55	Janoos	@newshour	Negative
t_8	2:57	Henpow2nd	@drthamasho	negative

4. ARCHITECTURE

Initially the tweets are loaded into the database. Later the Sentiment classification (SC) model is applied for the collected data stream. On application of the model, we get to extract the sentiment classification of the data or the tweets. Once the sentiment of the tweets is extracted, the sentiments of the tweets are classified and its features are identified along with the variance for the same

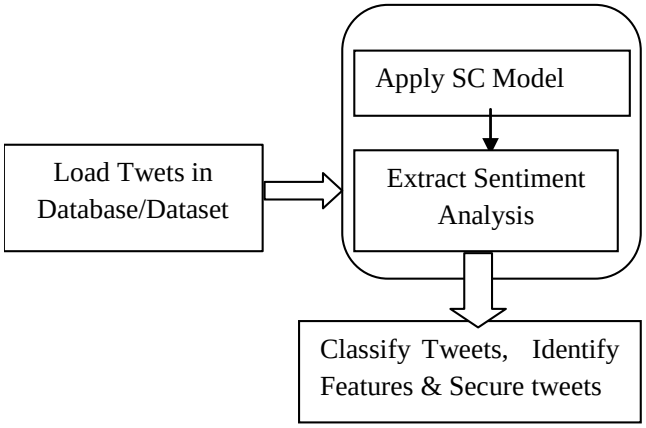

Fig. 2: Architecture

5. Features

The text feature set and non text feature are used simultaneously at client level and @ system based elements.

Text features: The sentiment words that are adaptive to topics and commonly used are taken. wordNet influence [24] furthermore, public feeling dictionary [25] are used to obtain the words. The google search engine is considered as the kernel and very huge querying hit is considered. PMI values are calculated from the same and then orientation words are separated. With labeling for tweets on a subject and evacuating the opinion words in common the successive descriptive words, verbs, things are extracted.

Non text features: Many non text features are considered here. The *Temporal Features* that is different from the traditional web documents. The users' views are associated well with their clock. So we extract a time post tweets as temporal tweets. *Emoticon Features* are collected from the Wikipedia as dictionary. Labelling is done for the emoticons as (+1) for positive, (0) for neutral and (-1) for negative. Corresponding values of the emoticons are summed up to its emoticons feature values. *Punctuation Features* are also a part of tweets that represent the users' sentiment. Such punctuations are also considered for the analysis. *User Level Tweets* considering the previous observations, we could predict the consistency in the users' tweets and in turn the prediction may be done on the opinion of the user. *@ Network Based Features* two alternative for the @network are there namely parent and child. For every parent and child node, the value is denoted and assigned. Through this the sentiment of parent node and child node is calculated differently and stored.

6. RESULTS

The sentiment classification model along with @ network based features gives more accurate result when compared to results of model without system based components. The exactness is increased by at least near to 16% and the f score is increased by at least near to 30%. The results with different step lengths and different sample ratio also proves that the model is quite reliable when compared to the current analysers without the @ values. The model is adaptive to different topics and thus the usability of the model increases and hence the efficiency of the model.

CONCLUSIONS

Different topics are discussed in twitter. Classification of sentiments on tweets suffers from lack of labelling of the tweet and adapting to the unpredictable topics. We formally propose a SVM model for training system. Contrasted and the surely understood baselines, model accomplishes increase in exactness that can be reliable.

REFERENCES

[1] B. J. Jansen, M. Zhang, K. Sobel, and A.Chowdury, "Twitter power: Tweets as electronic word of mouth," Journal of the American society for information science and technology, vol. 60, no. 11, pp.2169–2188, 2009.

[2] ——, "Micro-blogging as online word of mouth branding," in CHI'09 Extended Abstracts on Human Factors in Computing Systems. ACM, 2009, pp. 3859–3864.

[3] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," Journal of Computational Science, vol. 2, no. 1, pp. 1–8, 2011.

[4] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment." ICWSM, vol. 10, pp. 178–185, 2010.

[5] C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou, and P. Li, "User-level sentiment analysis incorporating social networks," in Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '11. New York, NY, USA: ACM, 2011, pp. 1397–1405.

[6] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification," in ACL, vol. 7, 2007, pp. 440–447.

[7] F. Li, S. J. Pan, O. Jin, Q. Yang, and X. Zhu, "Cross-domain co- extraction of sentiment and topic lexicons," in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. Association for Computational Linguistics, 2012, pp. 410–419.

[8] S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen, "Cross-domain sentiment classification via spectral feature alignment," in Proceedings of the 19th international conference on World wide web. ACM, 2010, pp. 751–760.

[9] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," CS224N Project Report, Stanford, pp. 1–12, 2009.

[10] S. Li, C.-R. Huang, G. Zhou, and S. Y. M. Lee, "Employing personal/impersonal views in supervised and semi-supervised sentiment classification," in Proceedings of the 48th annual meeting of the association for computational linguistics. Association for Computational Linguistics, 2010, pp. 414–423.
[11] S. Li, Z. Wang, G. Zhou, and S. Y. M. Lee, "Semi-supervised learning for imbalanced sentiment classification," in Proceedings of the Twenty-Second international joint conference on Artificial Intelligence- Volume Volume Three. AAAI Press, 2011, pp. 1826–1831.
[12] X. Wan, "Co-training for cross-lingual sentiment classification," in Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1. Association for Computational Linguistics, 2009, pp. 235–243.
[13] N. Yu and S. K¨ubler, "Filling the gap: Semi-supervised learning for opinion detection across domains," in Proceedings of the Fifteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, 2011, pp. 200–209.
[14] S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen, "Cross-domain sentiment classification via spectral feature alignment," in Proceedings of the 19th international conference on World wide web. ACM, 2010, pp. 751–760.
[15] F. Li, S. Wang, S. Liu, and M. Zhang, "Suit: A supervised useritem based topic model for sentiment analysis," in Proceedings of Twenty-Eighth AAAI Conference on Artificial Intelligence, ser. AAAI- 14, 2014, pp. 1636–1642.
[16] S. Liu, F. Li, F. Li, X. Cheng, and H. Shen, "Adaptive co-training svm for sentiment classification on tweets," in Proceedings of the 22Nd ACM International Conference on Conference on Information & Knowledge Management, ser. CIKM '13. New York, NY, USA: ACM, 2013, pp. 2079–2088. [Online]. Available: http://doi.acm.org/10.1145/2505515.2505569
[17] K.-L. Liu, W.-J. Li, and M. Guo, "Emoticon smoothed language models for twitter sentiment analysis." in AAAI, 2012.
[18] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of twitter data," in Proceedings of the Workshop on Languages in Social Media. Association for Computational Linguistics, 2011, pp. 30–38.
[19] S. Liu, W. Zhu, N. Xu, F. Li, X.-q. Cheng, Y. Liu, and Y. Wang, "Cotraining and visualizing sentiment evolvement for tweet events," in Proceedings of the 22nd international conference on World Wide Web companion. International World Wide Web Conferences Steering Committee, 2013, pp. 105–106.
[20] S. Havre, E. G. Hetzler, and L. T. Nowell, "ThemeRiver: Visualizing Theme Changes over Time," in IEEE Symposium on InformatioVisualization, 2000, pp. 115–124.
[21] S. L. Havre, E. G. Hetzler, P. D. Whitney, and L. T. Nowell, "ThemeRiver: Visualizing Thematic Changes in Large Document Collections," IEEE Transactions on Visualization and Computer Graphics, vol. 8, pp. 9–20, 2002.
[22] L. Byron and M. Wattenberg, "Stacked Graphs - Geometry & Aes- thetics," IEEE Transactions on Visualization and Computer Graphics,vol. 14, pp. 1245–1252, 2008.