Sentiment Analysis of Social web data: A Review

[1]Raj Kumar Verma, [2]Dr. Ritu Tiwari, [3]Prof. Nirmal Roberts

Robotics and Intelligent System Lab, ABV-IIITM Gwalior, India

***Abstract*—Today social networking websites has evolved to become a source of various kind of information. This is because of the nature of these websites on which peoples comments and post their opinions on different types of topics i.e. they express positive or negative sentiments about any product that they use in daily life, complains and current issues etc. These sentiments help in getting information about various current trends and can be used further in deciding usefulness of some tasks, products and themes. Also social web data like twitter has a large amount of data that people post so it's become important to work on efficient intelligent systems that can do data refinement, analysis of tasks intelligently and efficiently. This paper presents a comprehensive overview of past and current research on twitter sentiment analysis and identifies outstanding research questions for the future.**

***Keywords*— TASC (Topic Adaptive Sentiment Classification), Query term, RA (Rated Aspect), ESLAM (Emoticon Smoothed Language Model), LARA (Latent Aspect Rating Analysis)**

I. Introduction

SENTIMENT ANALYSIS refers to the use of natural language processing, text and speech analysis and computational linguistics to identify and extract subjective information in source materials. It aims to determine the useful information out of the bulk of data and that information can be used to make some facts or predictions [20].

There are a lot of social websites that provide whole bulk of data that is informative from various view points. Twitter is one of the main source of such data. Twitter sentiments help in getting information about various current trends and can be used further in deciding usefulness of some tasks, products and themes. This data may have several categories like sports, food, person, awards, weather etc. Sentiment analysis on such data classify the polarity [7] of a given tweet at the document, sentence, or feature/aspect level. As web data like twitter has a large amount of data that people post so it's become important to work on efficient intelligent systems that can do data refinement, analysis of tasks intelligently and efficiently. A significant research work has been done since 1980 on different aspects of twitter sentiment analysis. Now we will see major work in this field and their contributions.

II. Past research work

A. *Sentence based analysis [1]*

In this paper author studies text to speech analysis of data. They studies about processing of data and ability to render natural expressive speech. The work basically highlights the techniques that are used in natural expressive speech and how this data is being used. Next we will see the process of natural expressive speech using the diagram shown by author. The diagram explains about input data processing and how data is being classified.

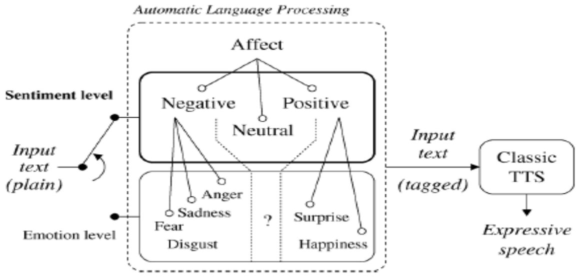


Fig 1 Framework of the proposed expressive TTS system for automatically detecting textual affect-related information represented in the hierarchy of affect [1]

This work basically focuses on production of synthetic speech and adaption of sentiment analysis[20] procedure that can then be used as input feature for expressive speech synthesis. So the main work of this paper revolves around data to speech synthesis. From the fig 1 we can see that data is classified into different categories of emotions and given to expressive test to speech.

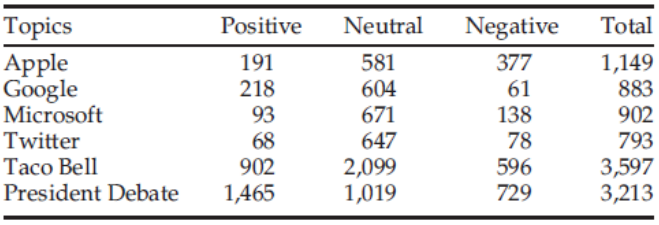
This paper mainly concerned about the identification of different effects of text and web data and tries to classify data into different effects. This work was the first attempt to adapt conventional sentiment analysis work. In this work author has used only small training data so the classification can be improved in terms of speed and accuracy if large data sets are used. Using a fast scripting language the above functionalities can be achieved up to a level.

B. *Field adaptive classification of tweets[2]*

In this paper author studies classifications of tweets in field adaptive manner. As topics in twitter are very diverse, it is impossible to train a universal classifier for all topics[17] and twitter lacks data labeling so its becomes more difficult to make them classify into different fields. The TASC algorithm that the author uses updates topic-adaptive features based on the collaborative selection of unlabeled data , which in turns helps to select more reliable tweets to boost the performance.

The algorithm that the author uses for classification of tweets into different feature fields is shown in above table 1. The above table shows results of sample data showing total tweets and their behavior in different categories. They also design the adapting model along a timeline (TASC-t) for

Table I
CORPUS STATISTICS[2]



Topics	Positive	Neutral	Negative	Total
Apple	191	581	377	1,149
Google	218	604	61	883
Microsoft	93	671	138	902
Twitter	68	647	78	793
Taco Bell	902	2,099	596	3,597
President Debate	1,465	1,019	729	3,213

dynamic tweets and experimented on 6 topics from published tweet corpuses demonstrating that usefulness of the TASC algorithm.Their future work emphasized on accurate labeling of data and a good classifier to model dynamic tweets[20] into feature field.

.

C. *User level sentiment analysis[3]*

In this paper author emphasizes on social relationships to improve user level sentiment analysis. Author’s approach is that users which are connected are more likely to hold similar opinions therefore relationship information can complement what we can extract about a users view points from their utterances. Propose models that are induced either from the Twitter follower/followee network or from the network in Twitter formed by users referring to each other using “@” mentions.

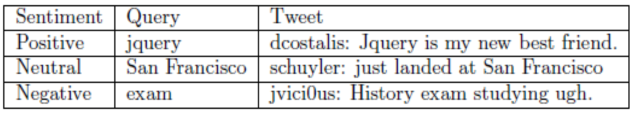
Results reveal that incorporating social-network information can indeed lead to statistically significant sentiment classification improvements over the performance of an approach based on Support Vector Machines having access only to textual features.

So the work mainly emphasizes on social relationship of users on web and then using that information to do user level analysis of data. This paper explore social network structures to help sentiment analysis[18], represents an interesting research direction in social network mining. The future work is to build more labeled targets sets. Also, datasets from other online social media systems with other kinds of social networks and more information on users would also be worth exploring.

D. *Sentiment classification using distant vision[4]*

In this paper author gives an approach to classify the twitter data based on query terms thus gives a new approach. This is basically useful for the customers who use sentiments before buying something and for the companies who keep track of their brand sentiments reviews. In this author shows good accuracy of classification using machine learning algorithms and by use of distant vision.

TABLE II
EXAMPLE TWEETS[4]



Sentiment	Query	Tweet
Positive	jquery	dcostalis: Jquery is my new best friend.
Neutral	San Francisco	schuyler: just landed at San Francisco
Negative	exam	jvici0us: History exam studying ugh.

Above table 2 shows the examples tweets that are being used as sample test and query term that is being used for the classification into different behaviors. The main idea of this work is using tweets with emoticons for distant supervised learning. The future work emphasizes on improving the accuracy of the classification. If we reduce the domains then the accuracy can be still improved greatly. Internationalization can be applied instead of just English sentences as twitter[20] has people from all over the world.

E. *RA summarization of short comments [5]*

In this paper author proposes work to summarize the comments for some product and item based on user votes. Basically it is the study of generating a RA summary of short comments, which is a decomposed view of the overall ratings for the major aspects so that a user could gain different perspectives towards the target entity.

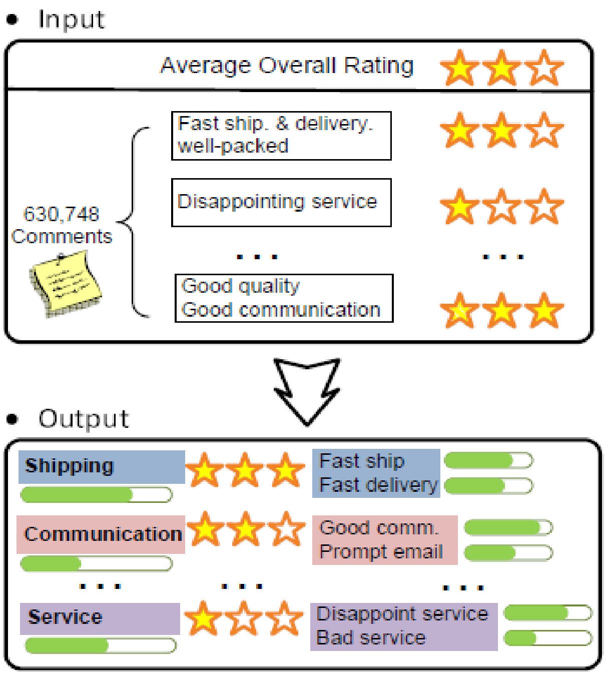


Fig 2 Problem Setup[5]

In the fig 2 input data represents what users normally can see through a website of community with comments which usually consist of large number of comments with companion overall ratings. In output overall rating is decomposed into several aspects each having support information showing the confidence on the aspect rating.

The future work mainly emphasize on the comparative study of product reviews and seller services. This rated mechanism can be applied with product reviews summarization to produce more accuracy in generated output text.

F. *Joint Model of Text and Aspect Rating[6]*

In this paper author proposes a statistical model find topics in text and extract textual evidence from reviews supporting each of these aspect ratings – a fundamental problem in aspect-based sentiment summarization. The accuracy achieved by the model is good. In this author also incorporated the aspect based summarization of text and rating. The approach which is suggested by the author is very general and can be easily used for the segmentation of other application where data is sequential with correlated signals.

The future work emphasize on incorporating the model into an end to end sentiment summarization system in order to evaluate at that level.

G. *Other related work*

Other research work includes cross domain sentiment classification [7] .Using this feature augmentation and selection according to the information gain criteria for the cross domain sentiment classification work performed better compared to other approaches. Sentiment work is also done on some real time themes like sentiment performance in characterizing debate performance [8]. They have shown that visuals and metrics can be used to inform the design of visual analytic systems for social media events. Such sections can identify key sections of a debate performance.

Sentiment classification is also done on microblogs [9]. Microblogs as a new textual domain offer a unique proposition for sentiment analysis. Their short document length suggests any sentiment they contain is compact and explicit. In some way classifying sentiments in microblogs[17] is easier then blogs and make a number of observations pertaining to the challenge of SL for sentiment analysis on microblogs. Apart from this work is also done in emoticons areas of twitter [10]. In this emoticons smoothed language models are build for the data cleaning and classification. This model is known as ESLAM. Extraction of emoticons [11] is another area in sentiment analysis category where a lot of work is done to extract different behavior parts of a tweet and text. For this supervised multi-engine classifier approach is used to identify emotion topic(s) from English blog sentences.

Other work includes Highlighting Disputed Claims on the Web[12] . It includes building a Dispute Finder, a browser extension that alerts a user when information they read online is disputed by a source that they might trust. Other work includes latent aspect rating analysis [13]. In this a new opinionated text data analysis problem called LARA, which aims at analyzing opinions expressed about an entity in an online review at the level of topical aspects to find each individual reviewer's latent opinion on each aspect as well as the relative emphasis on different aspects when forming the overall judgment of the entity. Other majors studies are done on mining of web data and summarization techniques [14]. As we know that summarization is one of the hardest problem of data mining and to build a system that can handle this work efficiently is a challenging task. Other work includes prediction of collective sentiment dynamics from time series social media data [15]. Predictive analysis allows the stakeholders to leverage immediate, accessible and vast reachable communication channel to proact and react against the public opinion. Apart from twitter data also analyzed for the prediction tasks like elections, whether etc [16].

So in this way we can see that a lot of study is done on different uses of sentiments analysis. Twitter data so vast that the new study areas are emerging each day. Still there are areas in existing research work where improvements can be done. We will discuss these improvements and tasks . In the next section we are going to discuss all improvements and work that can be done to improve the data analysis work so that more refinements can be done in social web data.

III. Discussion

We have seen that a lot of work is done in sentiment analysis field using any social website like twitter and many techniques are devised to improve accuracy of classification of social sentiments. The future work gives us idea about improving the classification and accuracy of social data. Mainly good optimizations can be done in classification part using good classifiers. Topic modeling is one area where limited work is done and also it is not applied at a big scale on social data like twitter sentiments. Topic modeling is one area which can help to divide large social data into categories by building an intelligent system. Apart from this summarization of data is another field which can be explored at a bigger level. Both topic modeling and summarization can be applied together to generate a intelligent system that can be useful in giving useful information out of a huge bulk of data from social web.

IV. Conclusion

As social web like twitter is so vast that getting all information is almost infeasible but possible steps can be taken to get most of the useful information from it. This can be achieved through sentiment analysis and there are various areas in sentiment analysis field like data refinement, topic modeling of sentiments and summarization of sentiment tweets which are still untouched and there are strong chances that we can get a lot of new outcomes ,techniques and methods if we explorer these areas.

References

[1] Alexandre Trilla and Francesc Alías;"Sentence-Based Sentiment Analysis for Expressive Text-to-Speech" IEEE transaction on audio, speech and language processing, ,February 2013, Page 223-233

[2] Shenghua Liu, Xueqi Cheng, Fuxin Li, and Fangtao Li;"Topic-Adaptive Sentiment Classification on Dynamic Tweets" IEEE Trans. Knowl. Data Eng. 27(6): 1696-1709 (2015)

[3] C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou, and P. Li, "User-level sentiment analysis incorporating social networks," in Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2011, pp. 1397–1405

[4] 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.

[5] Yue Lu, ChengXiang Zhai, Neel Sundaresan; "Rated aspect summarization of short comments" In www 2009, pp 131-140

[6] Ivan Titov, Ryan Mcdonald; "A Joint Model of Text and Aspect Ratings for Sentiment Summarization" In Proceedings of ACL-08: HLT (June 2008), pp. 308-316

[7] Y. He, C. Lin, and H. Alani, "Automatically extracting polarity bearing topics for cross-domain sentiment classification," in Proc. 49th Annu. Meeting Assoc. Comput. Linguistics: Human Language Technol.-Volume 1, 2011, pp. 123–131.

[8] N. A. Diakopoulos and D. A. Shamma, "Characterizing debate performance via aggregated twitter sentiment," in Proc. SIGCHI Conf. Human Factors Comput. Syst., 2010, pp. 1195–1198.

[9] Adam Bermingham & Alan Smeaton; " Classifying Sentiment in Microblogs: Is Brevity an Advantage?" ACM, New York, NY, USA, 1833-1836.2010

[10] K.-L. Liu, W.-J. Li, and M. Guo, "Emoticon smoothed language models for twitter sentiment analysis." in Proc. 26th AAAI Conf. Artif. Intell., 2012, pp. 1678–1684.

[11] D. Das and S. Bandyopadhyay, "Extracting emotion topics from blog sentences: Use of voting from multi-engine supervised classifiers," in Proc. 2nd Int. Workshop Search Mining User-Generated Contents 19th ACM Int. Conf. Inform. Knowl. Manage., 2010, pp. 119–126.

[12] Rob Ennals, Beth Trushkowsky, John Mark Agosta, Tye Rattenbury, and Tad Hirsch ;" Highlighting Disputed Claims on the Web" ACM International World Wide Web Conference (WWW), 2010

[13] Hongning Wang, Yue Lu, Chengxiang Zhai;" Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach" ACM New York, NY, USA 2010

[14] M. Hu and B. Liu, "Mining and summarizing customer reviews," in Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2004, pp. 168–177

[15] 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.

[16] 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

[17] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment in twitterevents," J. Am. Soc. Inform. Sci. Technol., vol. 62, no. 2, pp. 406–418,2011.

[18] X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang, "Topic sentiment analysis in twitter: A graph-based hashtag sentiment classification approach," in Proc. 20th ACM Int. Conf. Inform. Knowl. Manage., 2011, pp. 1031–1040.

[19] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter power: Tweets as electronic word of mouth," J. Am. Soc. Inform. Sci. Technol., vol. 60, no. 11, pp. 2169–2188, 2009.

[20] 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.