Sentiment Analysis for Social Media: A Survey

Harshali P. Patil
Department of Computer Engineering
Thakur College of Engineering & Technology
Mumbai, India
harshali.patil9@gmail.com

Dr. Mohammad Atique
Department of Computer Science & Engineering
SGBAU
Amravati, India
mohd.atique@gmail.com

Abstract*—In the past years, the World Wide Web (WWW) has become a huge source of user-generated content and opinionative data. Using social media, such as Twitter,facebook,etc,user share their views,feelings in a convenient way. Social media, such as Twitter,facebook,etc, where millions of people express their views in their daily interaction,which can be their sentiments and opinions about particular thing. These ever-growing subjective data are, undoubtedly, an extremely rich source of information for any kind of decision making process. To automate the analysis of such data, the area of Sentiment Analysis has emerged. It aims at identifying opinionative data in the Web and classifying them according to their polarity,i.e., whether they carry a positive or negative connotation. Sentiment Analysis is a problem of text based analysis, but there are some challenges that make it difficult as compared to traditional text based analysis This clearly states that there is need of an attempt to work towards these problems and it has opened up several opportunities for future research for handling negations, hidden sentiments identification, slangs,polysemy. However, the growing scale of data demands automatic data analysis techniques. In this paper, a detailed survey on different techniques used in Sentiment Analysis is carried out to understand the level of work.

Keywords—sentiment analysis; opinion;polarity;corpus.

I. INTRODUCTION

Sentiment analysis which is also known as opinion mining, is a method to automatic finding of opinions incarnated in text, is becoming a challenge in many research areas, particularly in data mining field for social media with a number of applications including product ratings and feedback analysis and customer decision making etc [9]. Currently social media has become a major public opinion finder and dissemination platform. With the expeditious development of Web 2.0, more and more people like to express their thoughts, views and approach over Internet, which increase the vast source of user-generated content and opinionative data.

The contribution of this survey is useful for many reasons. First, the survey provides cataloging of a large number of recent articles according to the techniques used. This approach could help out the researchers who want to use certain techniques in the Sentiment Analysis field and select the appropriate technique for a certain application. Finally, the survey gives some focus on some challenges related fields to sentiment analysis.

This paper is organized as follows: Section 2 tackles the different approaches for sentiment analysis and their related articles. Some challenges related to sentiment analysis fields are presented in Section 3. Finally, we conclude this paper in Section 4.

II. APPROACHES FOR SENTIMENT ANALYSIS

Authors of [1] formally stated their task, and interpreted how to mathematically incorporate social context and topical context into the basic prediction model. They investigated the content-based correlations among the topics, and calculated TCS to measure them. The assumptions about social context and topical context were both corroborated by the hypothesis testing over the Twitter data set they created. Finally, they conducted experiments to evaluate the proposed ScTcMF framework, and the experimental results demonstrated that both social context and topical context can help to improve the performance for the user topic opinion prediction.

According to [2] in sentiment analysis, the performance of Bag of words sometimes remains limited due to some fundamental deficiencies in handling the polarity shift problem.So,to address this problem for sentiment classification they proposed a model called dual sentiment analysis (DSA). They first proposed a novel data expansion technique by creating a sentiment-reversed review for each training and test review. On this basis, they proposed a dual training algorithm to make use of original and reversed training reviews in pairs for learning a sentiment classifier, and a dual prediction algorithm to categorize the test reviews by considering two sides of one review. They also extended the DSA framework from polarity (positive-negative) classification to 3-class (positive -negative-neutral) classification, by taking the neutral feelings into consideration. Finally, they developed a corpus- method to construct a pseudo-antonym dictionary.

Authors of [3] concentrated on engineering students' Twitter posts to understand problems and glitches in their educational experiences. They first conducted a qualitative analysis on samples taken from about 25,000 *tweets* associated to engineering students' college life. They found engineering students encounter problems such as deficiency of social engagement, heavy study load, and sleep deficiency. Based on these results, we implemented a multi-label classification algorithm to categorize tweets reflecting students' problems. They then used the algorithm to train a detector of student problems from about 35,000 tweets issued at the geo-location of Purdue University. This work, presents an approach and

978-1-4673-8611-1/15/$31.00 ©2015 IEEE

results that show how casual social media data can provide insights into students' experiences.

For the classification of sentiment applying sentiment classifier trained results in poor performance because words that occur in the train (source) domain might not appear in the test (target) domain so, To overcome the feature mismatch problem in cross-domain sentiment classification [4] proposed a cross-domain sentiment classifier using an automatically extracted sentiment sensitive thesaurus. They have done the comparisons against the SentiWordNet, a lexical resource for word polarity. They show that the created sentiment-sensitive thesaurus accurately captures words that express similar sentiments [4].

Authors observed that previous research mainly focused on modeling and tracking public sentiment so, they moved one step further to interpret sentiment variations. They worked on twitter dataset. They observed that emerging topics (named foreground topics) within the sentiment variation periods are highly related to the actual reasons behind the variations. These foreground topics can help to interpret the sentiment variations .Based on this observation, they proposed a Latent Dirichlet Allocation (LDA) based model, Foreground and Background LDA (FB-LDA), to dig out foreground topics and filter out background topics. To further improve the readability of the mined reasons, they ranked them with respect to their "popularity" within the variation period using Reason Candidate and Background LDA (RCB-LDA) method [5].

Authors presented a joint sentiment-topic model and a reparameterized version of JST called Reverse-JST when most of the existing approaches to sentiment classification favor supervised learning. Unlike supervised approaches to sentiment classification which often fail to produce acceptable performance when shifting to new domains, the weakly supervised environment of JST makes it highly convenient to other domains [6].

Investigation of [7] illustrates that three types of information is useful to automatically generate the overall sentiment polarity for a given hash tag in a definite time period, which markedly varies from the conventional sentence-level and document-level sentiment analysis, including sentiment polarity of tweets containing the hash tag, hash tags co-occurrence relation-ship and the literal meaning of hash tags in order to incorporate the first two types of information into a classification framework where hashtags can be classified collectively, they propose a novel graph model and investigate three approximate collective classification algorithms for inference. Going one step ahead, they show that the performance can be remarkably improved using an enhanced boosting classification setting in which we employ the literal meaning of hash tags as semi-supervised information. They worked on a real-life data set consisting of 29195 tweets and 2181 hash tags to show the effectiveness of their proposed model and algorithms.

Authors of [8] have shown that the sentiment analysis results produced by their hybrid approach are favorable compared to the lexicon-only and learning-only baselines. For both sentiment polarity classification and sentiment strength detection, their pSenti system, achieves high accuracy that is very close to the pure learning-based system, and much higher than the pure lexicon-based system .This approach is able to combine the best of two worlds | the stability as well as readability from a carefully designed lexicon, and the high accuracy from a powerful supervised learning algorithm.

Authors have explored the predictive power of reviews using the movie domain as a case study, and studied the problem of predicting sales performance using sentiment information mined from reviews [9]. They propose Sentiment PLSA (S-PLSA), in which a review is considered as a document generated by a number of hidden sentiment factors, in order to capture the complex nature of sentiments. Then they propose ARSA, an Autoregressive Sentiment-Aware model for sales prediction. Then they search for further improvement in the accuracy of prediction by considering the quality factor, with a focus on predicting the quality of a review in the absence of user-supplied indicators, and present ARSQA, an Autoregressive Sentiment and Quality Aware model, to use sentiments and quality for predicting product sales performance.

It is possible for the stock price of some companies to be predicted with an average accuracy as high as 76.12%.They proposed a method to mine Twitter data for answers to the questions like if the price of a selection of 30 companies listed in NASDAQ and the New York Stock Exchange can actually be predicted by the given 15 million records of Twitter messages [10]. We have summarized the survey in Table I for the different type of work done in the Sentiment Analysis field. R* is the reference number. The reason behind selecting the columns of the table is to just to analyze the work done in the sentiment analysis field.

III. CHALLENGES

A. Incremental Approach

Analysis of real time data is not one time operation. Whenever data is added we need to do analysis then why should not we use the previous analysis result. Incremental approach allows an existing result to be updated using only new individual data instances, without having to re-process past instances. This may be useful in situations where the entire dataset is not available when the data changes over time.

B. Parallel Computing For Massive Data

If we divide the computation into tasks or processes that can be executed simultaneously, then there can be an improvement

TABLE I: TECHNIQUES FOR SENTIMENT ANALYSIS

R*	Approach	Tools/Techniques	Experiment	Language Dependency	M/c Learning/ Lexicon Based(ML*/LB*)	Data Scope	Data Source
1	User-Topic opinion prediction (2013)	Social context and Topical context incorporated Matrix Factorization (ScTcMF)	To predict the unknown user-topic opinions.	Yes	LB*	Twitter	Tweets
2	Polarity shift in sentiment classification(2015).	Dual sentiment analysis(DSA)	Polarity classification task	No	LB*	Multi-domain sentiment English dataset. two Chinese dataset	Amazon.com, ChnSentiCorp corpus
3	Qualitative analysis and large-scale data mining techniques(2014)	Naïve-Bayes multi-label classification algorithm	Show how informal social media data can provide insights into students' experiences.	Yes	ML*	Twitter	Tweets
4	Cross-domain sentiment classification(2013)	Corpus based	To evaluate the benefit of using a sentiment sensitive thesaurus for cross-domain sentiment classification	Yes	LB*	Product reviews	Amazon.com
5	To interpret sentiment variations (2014)	Latent Dirichlet Allocation (LDA) based model, Foreground and Background LDA (FB-LDA), generative model called Reason Candidate and Background LDA (RCB-LDA	To mine possible reasons of public sentiment Variations.	Yes	ML*	Twitter	Tweets
6	Sentiment and topic detection (2012)	Weakly supervised joint sentiment-topic (JST) model based on latent Dirichlet allocation (LDA, Reverse-JST	To detect sentiment and topic simultaneously from text	Yes	ML*	Product reviews, Movie reviews	Amazon.com, IMDB movie archive
7	Hashtag-level sentiment classification(2011)	SVM classifier	To automatically generate the overall sentiment polarity for a given hashtag in a certain time period, which markedly differs from the conventional sentence-level and document-level sentiment analysis.	Yes	ML*	Self-annotation manner to label the dataset, Twitter	Tweets
8	Sentiment polarity classification and sentiment strength detection (2012)	Hybrid approach(lexicon based +M/c learning)	To classify polarity and detect sentiment strength	Yes	ML* and LB*	Software reviews and movie reviews	CNET, IMDB
9	Sales prediction (2012)	Sentiment PLSA (S-PLSA, ARSQA, an Autoregressive Sentiment and Quality Aware model	To Predict Sales Performance	Yes	ML*	Movie reviews	IMDB
10	Predicting Stock Price Movements (2014)	NLP techniques	To determine if the price of a selection of 30 companies listed in NASDAQ and the New York Stock Exchange can actually be predicted by the given 15 million records of tweets	Yes	ML*	Twitter	Tweets

in the speed through the use of parallelism, it is necessary to achieve this in sentiment analysis for massive data of social media, where massive instant messages are published every day so that we can utilize the overall computing power.

C. Credibility/Behavior/Homophily

Behaviors in social media are only observed by the traces they leave in social media. We rarely observe the driving factors that cause these behaviors; nor can we interview individuals regarding their behaviors. Even if a behavior is analyzed on social media and related patterns are gleaned, it's difficult to verify the validity of these behavioral patterns. Evaluation becomes even more challenging for industries in which important decisions are to be made based on observations of individual behavior.

D. Sarcasm

Sarcasm can be used to hurt or offend or can be used for comic affect.It means false positives for eg."Children really brighten up a household - they never turn the lights off". Detecting sarcasm from the expressions and finding out the correct context related sentiments is a challenging task. It is an ironic or satirical remark that seems to be praising someone or something but is really taunting or cutting.

E. Grammatically Incorrect Words

There are many approaches that analyze sentiments but hardly any work accomplished on grammatical errors. The results of sentiment analysis can be improved if these types of errors can be mapped to correct words.

F. Review Author Segmentation

Opinion towards a target may be specified by many people who can be called as review authors. Depending on the commenting style of these authors, they should be categorized so that credibility evaluation will be easy. In decision making this credibility evaluation is helpful.

G. Refinement Of existing Lexicons or Updating/Down-Dating Lexicons

Many people comments, the Performance of sentiment analyzer depend on the correctness of the lexicon. Fine-tuning of existing lexicons is required to accommodate new words and destroy the words which are no more used for better results. Lexicon expansion through the use of synonyms has a drawback of the wording loosing it primary meaning after a few recapitulation.

H. Handling Noise and Dynamism

Social media data are enormous, noisy, unstructured, and dynamic in nature, and thus novel challenges arise, introduces representative research problems of mining social media. Identifying and removal of noisy data is a challenging task.

IV. CONCLUSION

Masses of users share their feelings on social media, making it a valuable platform for tracking and exploring public sentiment. Social media is one of the biggest platforms where massive instant messages are published every day which makes it an ideal source for capturing the opinions towards various curious topics, such as products, goods or celebrities, etc. The main goal of this paper is to give an overview of latest updates in sentiment analysis and classification methods and it includes the brief discussion on the challenges of sentiment analysis for which the work needs to be done. We also found that most of the works done are based on machine learning method rather than the lexicon based method.

References

[1] Fuji Ren,Ye Wu," Predicting User-Topic Opinions in Twitter with Social and Topical Context", *IEEE* Trans. on affective computing, vol. 4, no. 4, October-December 2013.

[2] Rui Xia, Feng Xu, Chengqing Zong, Qianmu Li, Yong Qi, and Tao Li,"Dual Sentiment Analysis:Considering Two Sides of One Review", *IEEE* Trans.on Knowledge and Data Engineering, 2015

[3] Xin Chen,Mihaela Vorvoreanu, and Krishna Madhavan," Mining Social Media Data for Understanding Students' Learning Experiences" *IEEE* trans. on learning technologies, vol. 7, no. 3, July-September 2014.

[4] Danushka Bollegala, David Weir, and John Carroll," Cross-Domain Sentiment Classification Using a Sentiment Sensitive Thesaurus", *IEEE* trans. on knowledge and data engineering, vol. 25, no. 8, August 2013.

[5] Shulong Tan, Yang Li, Huan Sun, Ziyu Guan, Xifeng Yan, Jiajun Bu, ,Chun Chen, and Xiaofei He," Interpreting the Public Sentiment Variations on Twitter", *IEEE* trans. on knowledge and data engineering, vol. 26, no. 5, May 2014.

[6] Chenghua Lin, Yulan He, Richard Everson, Member, IEEE, and Stefan Ru¨ger," Weakly Supervised Joint Sentiment-Topic Detection from Text", *IEEE* trans. on knowledge and data engineering, vol. 24, no. 6, June 2012.

[7] Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, Ming Zhang," Topic Sentiment Analysis in Twitter: A Graph-based Hashtag Sentiment Classification Approach", *ACM*, CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK, 2011.

[8] Andrius Mudinas, Dell Zhang, Mark Levene," Combining Lexicon and Learning based Approaches for Concept-Level Sentiment Analysis", WISDOM' 12, August 12, 2012, Beijing, China Copyright 2012, *ACM.*

[9] Xiaohui Yu,Yang Liu, Jimmy Xiangji Huang, and Aijun An,," Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain", *IEEE Trans.On* knowledge and data engineering, vol. 24, no. 4, April 2012.

[10] LI Bing, Keith C.C. Chan, Carol OU," Public Sentiment Analysis in Twitter Data for Prediction of A Company's Stock Price Movements", 2014 *IEEE*, 11th International Conference on e-Business Engineering.

[11] Walaa Medhat , Ahmed Hassan , Hoda Korashy," Sentiment Analysis Algorithms and Applications: A survey", Ain Shams Engineering Journal (2014).