A Survey on Sentiment Analysis and Opinion Mining

Pooja C. Sangvikar
Department of Information Technology
Smt.Kashibai Navale College of Engineering Pune, India

Abstract— In Today's world, the social media has given web users a place for expressing and sharing their thoughts and opinions on different topics or events. For this purpose, the opinion mining has gained the importance. Sentiment classification and Opinion Mining is the study of people's opinion, emotions, attitude towards the product, services, etc. Sentiment Analysis and Opinion Mining are the two interchangeable terms. There are various approaches and techniques exist for Sentiment Analysis like Naïve Bayes, Decision Trees, Support Vector Machines, Random Forests, Maximum Entropy, etc. Opinion mining is a useful and beneficial way to scientific surveys, political polls, market research and business intelligence, etc. This paper presents a literature review of various techniques used for opinion mining and sentiment analysis.

Key words: Sentiment Analysis, Opinion Mining, Sentiment Classification

I. INTRODUCTION

Data Mining is one of the important step of the "Knowledge Discovery in Databases" processes or KDD, which discovers patterns from large datasets. Also, it includes the techniques and methods of artificial intelligence, machine learning and statistics. To extract the valuable information from a dataset and transform it into an understandable and simple structure, is the main purpose of data mining process.

Opinion mining or sentiment analysis is to analyze the users' opinions or thoughts which are in the form of unstructured data. The interpretation of meanings of views is a terrible task, for that Sentiment Analysis plays a vital role. To interpret and understand the person's views, emotions and thoughts, the system must be made reliable and efficient.

There are two techniques used in Sentiment Classification.

- Machine Learning Approach
- Lexicon Based Approach

Machine Learning techniques include supervised and unsupervised learning approaches. Supervised learning consists of some classifiers like Decision tree classifiers, Liner classifiers, Rule-based classifiers and Probabilistic classifiers. SVM and Neural Networks are the linear classifiers whereas Naïve Bayes, Bayesian Network, and Maximum Entropy is Probabilistic classifiers. Lexicon Based approaches are classified into Dictionary based and Corpus-based methods. The corpus-based approach further divided into the statistical and semantic approach.

The analysis of sentiments can be Document based where the opinion in the entire document is summarized or classified as positive, negative or neutral, other is Sentence based where various sentences in the text are classified or analyzed and Lastly, Sentiment Analysis can also be Phrase level based where the polarity is used, and classification is done. Various applications uses the Sentiment Analysis like in consumer market for product reviews, social media for finding general opinions about recent topics, to know the consumer trends etc.

The remainder of this paper is organized as follows: Section 2 gives an overview of related work already done on Sentiment Analysis. Section 3 includes result and a comparative table of existing systems. Section 4 concludes survey.

II. RELATED WORK

Sentiment Analysis has been done using different techniques or methods. Some works extract the meaning of the text, document or sentence level while others obtain connections between users to assign sentiment polarity for sentiment analysis. In particular, sentiment analysis of tweets has been done not only using approaches based on text, i.e. lexicon based classifiers, but also by combining Natural Language Processing and Machine Learning techniques.

Tang and Qin in [1] proposed A Joint Segmentation and Classification Framework for Sentence Level Classification. Here a candidate generation model is developed to produce segmentation candidates of a sentence and a segmentation ranking model used to score the usefulness of segmentation candidates for sentiment classification and also classification model for predicting the sentiment polarity. The experiments are performed on two datasets: a tweet dataset and a movie review dataset. The effectiveness of this approach is verified by applying it to these two datasets.

Neethu and Rajasri in [2], analyzed the Twitter posts about electronic products like mobiles, laptops, etc. using Machine Learning approach. A new feature vector is developed for classifying the tweets as positive, negative and extracts peoples' opinion about products. Classifiers like Nave Bayes, SVM, Maximum Entropy and Ensemble classifiers are used for classification. All these different classifiers have the same accuracy for the new feature vector. This feature vector performs well for electronic products.

Sentiment Analysis of Twitter has become one of the most important research Topic. Lek and Danny in [3] proposed an aspect-based sentiment classification approach to analyzing opinions for tweets. Different experiments are performed on the aspect-based sentiment classifier, to improve the performance of existing classifiers. The experimental results show that aspect level classifier approach outperforms existing methods. From experimental results, it suggests that a layered classification approach that uses the aspect-based classifier as the first layer classification and the tweet-level classifier as the second layer classification is more efficient than a classifier trained using target-dependent features. Also, this approach consistently improves the performance of existing sentiment classifiers.

Xia, Xu in [4] has proposed a model called Dual Sentiment Analysis (DSA). By creating sentiment reversed review for each training and test review, data expansion technique is proposed. Also, DSA framework is developed for sentiment classification. Multi-domain datasets are used for the experiment. Classification accuracy is compared with Naïve Bayes, Linear SVM, etc.

Topics in Twitter are more diverse and always dedicated themselves to a particular domain or subject. Liu, Li, and Cheng in [5], proposed an adaptive multiclass SVM model that transfers an initial common sentiment classifier to a topic adaptive. An iterative algorithm is proposed which is followed by three steps: optimization, unlabeled data selection and adaptive feature expansion steps. Comparison is done with the well-Known supervised sentiment classifiers like SVM, Decision Tree, Random Forest etc. and semi-supervised approaches. Accuracy of algorithm increased averagely 74% on the 6 topics from public tweet corpus.

There are different methods to analyze information or text from particular document. But to extract and analyze the valuable information from text messages from social sites, Batool, Khattak in [6] proposed the Precise tweet classification technique. To find keywords from messages, feature extraction is used. After that semantic based filtering of data is done to categorize the data into specific category. By doing filtering on data, the information gain is increased i.e. knowledge enhancer is used to avoid information loss.

To identify the interest of people’s or opinion towards social events, Zhou, Yong in [7] proposed a Tweet Sentiment Analysis Model (TSAM). Evaluation of TSAM is done on political dataset i.e. Australian Federal Election 2010. The proposed model provides fast and less expensive alternative approach over traditional polls for mining public views. The drawback of the approach is POS processing is not done. Also, Emotion Analysis is not included during Sentiment Analysis.

To analyze the data, one can use Support Vector Machine (SVM) which is supervised learning method, but Basari, Hussain in [8] proposed a hybrid method of SVM and Particle Swarm Optimization (PSO). To solve the problem of dual optimization, hybrid method is proposed. For extraction of more features, from prepared datasets, PSO is used. The accuracy of SVM is affected by the hybridization of SVM-PSO and which is 77%. The other variation of SVM can be used for improving the accuracy.

To perform Sentiment Analysis on News articles which gives fine grained analysis, Raina in [9] used Sentic Computing approach. For Concept-Level analysis, SenticNet; a publicly available resource is used. The performance of engine is measured by evaluation parameters like accuracy, precision, recall and F-measure. The proposed approach gives an accuracy 71.2%. This Method is only feasible for news articles i.e. it should be commercially viable. The drawback found is verb + noun pair is not included in Semantic Parser. So Semantic Parser can be improved by extracting verb + noun pair.

III. RESULT AND COMPARISON

By above related work, comparison of different mechanism used for the sentiment classification is done.

The Overview of various existing techniques, methods and algorithms are described in the following table. The Paper title, Datasets used, advantages, disadvantages and accuracy are included in the table.

	Paper title	Datasets Used	Techniques/ Algorithms	Advantages	Drawbacks	Accuracy
1)	A joint segmentation and classification framework for sentence level sentiment classification.	Tweet dataset and movie review dataset.	Training algorithm and prediction algorithm.	-A joint modelling framework outperforms over pipeline method in various experimental settings.	-In joint framework static datasets are used i.e. publicly available dataset.	81%.
2)	Sentiment analysis in twitter using machine learning techniques	Tweets of electronic products.	Naïve Bayes , SVM , Max entropy.	-Simpler & efficient by using machine learning techniques. -Performs well for electronic products.	- Domain dependent i.e. only feasible for electronic products.	NB-89.5% SVM, Max Entropy-90%
3)	Aspect based twitter sentiment Classification.	Stanford Twitter Sentiment (STS) consists of 1.6 million tweets.	Aspect based sentiment classification.	Aspect based classifier consistently increasing the performance of classifier.	-In the aspect based classification, Contextual information is not considered.	84%
4)	Dual Sentiment Analysis considering two sides of one review.	Chinese Hotel, Notebook, 4 English Dataset.	Dual Sentiment Analysis (DSA) model.	-Effective for polarity classification. - Strengthen the DSA algorithm by data expansion technique.	- In the Dual Sentiment Analysis, Complex polarity shift patterns are not used.	73%
5)	Adaptive Co-training SVM for Sentiment Classification of tweets.	Sanders Twitter Sentiment Corpus & presidential debate corpus.	Multiclass SVM, Adaptive co-training algorithm.	Adaptive co-training SVM expands topic adaptive feature.	-Less features are extracted. -Timeline visualization of result is not done.	74%

6) Precise Tweet Classification and Sentiment Analysis.	Diabetic relevant Dataset.	Precise Tweet Classification Model.	-This is a method to collect more precise information about specific Topic.	-	72.8%
7) Sentiment Analysis on Tweets for Social Events.	Australian Federal Election 2010.	Tweet Sentiment Analysis Model (TSAM).	Knowledge enhancer is used to avoid information loss.	-POS processing is not done. -Also, Emotion Analysis is not included during Sentiment Analysis.	-
8) Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization.	Movie review dataset.	Support Vector Machine and Particle Swarm Optimization (SVM-PSO).	-Method solves the problem of dual optimization. -Hybridization gives better result.	-To increase the accuracy, SVM with other optimization methods can be used.	77%
9) Sentiment Analysis in News Articles Using Sentic Computing.	Multi-Perspective Question Answering (MPQA) corpus, a large corpus of news articles.	News-focused opinion-mining engine, sentic computing-based Approach.	-Method is reliable For identifying neutral sentences.	-only feasible for news articles i.e. it should be commercially viable. -verb + noun pair is not included in Semantic Parser.	71.2%

Table 1: Overview of various existing techniques and algorithms

IV. Conclusion

Internet and Web technologies are continuously growing and expanding, so the space and scope in the area of information retrieval is also expanding. For that, Sentiment Analysis and Opinion Mining is considered as an interesting area of research due to its many applications in society. Above survey explore different techniques, also these Opinion Mining techniques suffers from a number of problems, such as accuracy, scalability, quality, standard of data, domain dependency, etc. These problems have to be tackled separately and those solutions can be used to improve the methods to do sentiment analysis and classification.

References

[1] Duyu Tang, Bing Qin, Furu Wei, Li Dong, Ting Liu, and Ming Zhou, "A joint Segmentation and Classification Framework For Sentence Level Classification", IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 23, NO. 11, NOVEMBER 2015.

[2] Neethu M S, Rajasree R, "Sentiment Analysis in Twitter using Machine Learning Techniques", 4th ICCCNT 2013 ,July 4 - 6, 2013, Tiruchengode, India.

[3] Hsiang Hui Lek and Danny C.C. Poo, "Aspect-based Twitter Sentiment Classification", 2013 IEEE 25th International Conference on Tools with Artificial Intelligence.

[4] Rui Xia, Feng Xu, Chengqing Zong, Qianmu Li, Yong Qi, and Tao Li, "Dual Sentiment Analysis: Considering Two Sides of One Review", IEEE Transactions on Knowledge and Data Engineering, MANUSCRIPT ID, 1041-4347 (c) 2015 IEEE.

[5] Shenghua Liu, Fuxin Li, Fangtao Li, Xueqi Cheng, Huawei Shen, "Adaptive Co-Training SVM for Sentiment Classification on Tweets", Copyright 2013 ACM 978-1-4503-2263-8/13/10.

[6] Rabia Batool, Asad Masood Khattak, Jahanzeb Maqbool and Sungyoung Lee, "Precise Tweet Classification and Sentiment Analysis", 978-1-4799-0174-6/13/$31.00 ©2013 IEEE.

[7] Xujuan Zhou, Xiaohui Tao, Jianming Yong, " Sentiment Analysis on Tweets for Social Events",2013 IEEE 17th International Conference on Computer Supported Cooperative Work in Design, 978-1-4673-6085-2/13/$31.00 ©2013 IEEE.

[8] Abd. Samad Hasan Basaria,*, Burairah Hussina, I. Gede Pramudya Anantaa, Junta Zeniarjab, "Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization", 1877-7058 © 2013 The Authors. Published by Elsevier Ltd.

[9] Prashant Raina, "Sentiment Analysis in News Articles Using Sentic Computing", 2013 IEEE 13th International Conference on Data Mining Workshops, © 2013 IEEE.