Review of Market Prediction Using Sentiment and Opinion Mining

Santosh Thakare **Dr. Sandeep R. Sirsat**

Abstract **- In this paper Sentiment analysis is an important research area. This paper studies online forums,hotspots detection and sentiment analysis and text mining approaches. And this paper aims to investigate the effect of media news on market performances. This paper also discuss various technique like machine learning,Natural language processing technique.**

Keywords:- Sentiment Analysis, Machine Learning, twitter.

I. INTRODUCTION

The welfare of modern societies depends on their market economies. Therefore, it is crucial to study markets and learn about their movements. The past decade has witnessed a dramatic change of the media landscape with digital social media channels (for ex. blogs, online forums, and social networking sites) forward of mouth supplementing media such as (Newspapers, magazines and television). Internet has fueled a fast growing market in personal opinions. Because Internet has enabled an increasing amount of user generated content (UGC) that potentially becomes primary sources of information for both consumers and businesses, and stock price predication remains an attractive topic for both researchers and investors.

In general, the market prediction is divided into technical or fundamental analysis. The analysis is based on their input data, historic market data to be used former and any other kind of information or news about the country, society, company etc.

Sentiment analysis, also called opinion mining is a type of natural language processing for tracking the mood of the public about a particular product or topic. It involves in building a system to collect and examine opinions about the product made in blogs posts, comments, reviews or tweets. Sentiment analysis can be useful in several ways. For ex, in marketing it helps in judging the success of an ad-campaign or new product launch. There are several challenges in sentiment analysis. the first is a opinion word that is considered to be positive in one situation may be considered negative in another situation. A second challenge is that people don't always express opinion in a same way. Most traditional text processing relies on the fact that small differences between two pieces of text don't change the meaning very much. In sentiment analysis" the picture was great " is very different from " the picture was not great " People can be contradictory in their statements. Most reviews will have both positive and negative comments, which is manageable by analyzing sentences one at a time. However ,in the more informal medium like twitter or blogs, the more likely people are to combine different opinions in the same sentence which is easy for a human to understand, but more difficult for a computer to parse. Sometimes even other people have difficulty understanding what someone thought based on short piece of text because it lacks context. for ex, "That movie was as good as its last movie" is entirely dependent on what the person expressing the opinion thought of previous model.

II. RELATED WORK

Until now investigates three streams of related work:

i) Dynamic Cluster analysis of online forums
ii) Sentiment analysis of web documents
iii) Web text mining using machine learning

i) Dynamic Cluster analysis of online forums:

Online forums are usually related to each other due to two reasons. First ,strong commonalities are shared by forums with similar topics or themes. For ex within an entertainment society the academy awards forum might be highly correlated to the golden Globes Award forum. Secondly emerging events will trigger a temporary correlation between certain forums. for ex the movie " No Country for old man" won "Best motion picture of the year" in the 2008 Academy awards. Which noticeable connection between the corresponding forums during Oscar season. Extensive research work has been conducted upon various types of interacting social networks such as dynamic networks upon individuals ,industrial manufacturers, listed companies, and online virtual communities. One important work from the Doctoral Thesis of Asavathiratham at MIT in 1996[1] created an influence model as tractable representation for the dynamics of networked. Markov chains. This work has been utilized by several scholars for ex[2] Where tools are developed automatically and unobtrusively learn the social network structure that arises within a human group based on wearable sensors.

Chose 662 main ceramic manufacturers in Guangdong Foshan ceramic industry cluster to construct a Competition Relationship Network and proved that the network defined by competition relationship is a highly

clustered scale-free network [3]. Correlated listed company network in stock markets constitutes another important research area in both Academia and Industry[4,5]. Regarding network dynamics of online virtual societies and communities [21].

It is observed that limited work is done to depict timely dynamics of online sports communities. Online sports forums within a virtual society are the focus of our study. Where machine learning is used to dynamically depict the interacting structure and to cluster the forums according to their emotional polarity.

ii) Sentiment analysis of web documents:

There are a variety of metrics to classify web documents including topics , structures, authors, time and so forth. Text classification based on its emotional polarity has become a newly emerged frontier appealing to the web mining community. To illustrate how it works Suppose you are considering a vacation in city " A", You might use a search engine online such as Google and shoot the query "A". It would be handy to know what fraction of the matches Google returns recommends "A" as a travel destination. Incorporating sentiment analysis into search engine and text retrieval technologies enables a more efficient and functional services for users [6]. Sentiment analysis has been utilized in applications such as news tracking and summarizing, online forums, file sharing, chatting rooms, blogging etc [7]. YouTube introduced sentiment classification technology early this year to categorize all its comments into "Poor" or "Good"[8] As a promising research area, text sentiment analysis has been extensively studied.[9, 10, 11-12,13,14] Where sentiment analysis is used for text classification tasks. Existing sentiment calculation approaches fall into two types: machine learning based approach and semantic orientation based approach. The language that have been studied include English, Chinese and Arabic.

iii) Web text mining using machine learning:-

To conduct clustering and forecasting of online forum hotspots, We use two machine learning approaches: K-means and SVM. K-means are used. These techniques has been studied and applied in a wide range of domains. For ex Bioinformatics [15,16,17] information security, pattern recognition[18,19,20] ,text classification. In addition, various derivatives of conventional K-means algorithm have been developed. Based on statistical learning, SVM is able to overcome problem such as over fitting and local minimum to achieve high generalization. Application of SVM includes text classification, image processing, and time series analysis. In our study machine learning is the key bridge between emotional polarity data and network dynamics.

III. DATA SETS

Most of the work in the field uses movie reviews data for classification.

Users opinion is a major factor for improvement of the quality of services rendered and enhancement of the deliverables. Blogs, review sites, data and micro blogs provide a good understanding of the reception level of the products and services.

i. Blogs :-

With an increasing usages of internet, blogging and blog pages are growing . Blogs pages have become most popular to express personal opinions. Bloggers records the daily events in their lives and express their opinions, feeling s and emotions in a blogs. Blogs are used as a source of opinion in many of the studies related to sentiment analysis**.**

ii. Review sites:-

For any user in making a purchasing decision. The opinions of others can be an important factor. A large and growing body of user-generated reviews is available on the internet. The reviewer's data used in most of the sentiment classification studies are collected from the e-commerce websites.

iii. Micro-blogging:-

Twitter is a popular micro-blogging service where users create status messages called "tweets". These tweets sometimes express opinions about different topics.

iv. Sentiment Classification:-

Much research exists on sentiment analysis of user opinion data, which mainly judges the polarities of user reviews. In this paper sentiment analysis is often conducted at one of the three levels: the document level, sentence level or attribute level. In relation to sentiment analysis , the literature indicate two types of techniques including machine learning and semantic orientation. In addition to that, the natural language processing techniques (NLP) is used in this area.

v. Machine Learning:-

The machine learning approaches applicable to sentiment analysis mostly belong to supervised classification. Thus it is called "supervised learning" in a machine learning based classification two sets of documents are required: training and test set. A training set is used by an automatic classifier to learn the differentiating characteristics of documents, and test set is used to validate the performance of the automatic classifier. A number of machine learning techniques have been adopted to classify the reviews.

Machine learning techniques like Naive bayes (NB), maximum entropy (ME), and support vector machine(SVM) have achieved great success in text categorization. The other most well known machine learning method in the natural language processing are K- Nearest neighbourhood, ID3,C5, centroid classifier, Window classifier and N-gram model. Naïve Bayes is simple but effective classification algorithm. The naïve bayes algorithm is widely used algorithm for document classification.

Support vector machine (SVM), a discriminative classifier, is considered the best text classification method(Rui Xia,2011). The support vector machine is a statistical classification method proposed by Vapnik(Kaiquan Xu,2011). Based on the structural risk minimization principle from the computational learning theory.

The idea behind the centroid classification algorithm is very simple and straightforward. Initially the prototype vector or centroid vector for each training class is calculated , then the similarity between a testing document to all centroid is computed.finaly based on these similarities, document is assigned to the class corresponding to the most similar centroid.

vi. Natural Language Processing Techniques:-

Natural Language Processing Techniques was developed to help people find business intelligence from free-form data; however, these methods lack strength in Detecting people's opinion[22], In the past decades, both industry and academia have been trying to find effective methods and tools to extract opinion-oriented information automatically from unstructured data[23]. Sentiment analysis has evolved from text mining (TM) and NLP,but aims to determine the sentiment of a speaker or writer with respect to some specific topics[24]. SA has greatly assisted decision makers in extracting opinions from un structured human-authored documents[23], which can be applied in various areas. It reduces the need for reading huge amount of documents to extract business opinions on a variety of topics.Thare are three main reasons to choose SA as a research approach.

I. It converts large unstructured content into a form that allows for specific predictions about particular outcomes, without institute market mechanisms.

II. It builds models to aggregate the opinions of the collective population and gains useful insights into group behavior to predict future trends.

III. It applies gathered information on how people react to particular objects and then design marketing and advertising campaigns.

IV. Evaluation & Discussion

The performance of different methods use for opinion mining is evaluated by calculating various metrics like precision, recall and F-measure. Precision is the fraction of retrieved instances that are relevant, while recall is the fraction of relevant instances that are retrieved. The two measures are sometimes used together in the F1 score (also F-Score or F-measure) is a measure of test's accuracy. A lot of work has been done on movie and product reviews. Movie review mining is a more challenging application than many other types of review mining. The challenges of movie review mining lie in that factual information is always mixed with real life review data and ironic words are used in writing movie reviews. Product review domain considerably differs from movie review domain because of two reasons.

Firstly ,there are specific comments in product reviews. People may like some features and dislike some others. Thus reviews consist of both positive and negative opinions, which make the task of classifying the review as positive or negative together.

V. Suggestions for Future Work

Market prediction mechanisms based on online text mining are just emerging to be investigated rigorously utilizing the radical peak of computational processing power and network speed in recent times. This research helps put into perspective the role of human reactions to events in the making of markets and can lead to a better understanding of market efficiencies and convergence via information absorption. In summary, this identifies the below areas or aspects in need of future research:

i) Semantics:- Many of the current work of research are still too much focused on word occurrence methods and they rarely even use wordNet. Moreover, semantic relations can be researched with different objectives, from defining weighting schemes for feature-representation to semantic compression or abstraction for feature reduction.

ii) Syntax:- Syntactic analysis techniques has received probably even less attention then semantic ones. More advanced Syntax based techniques like usages of parse trees for pattern recognition in text can improve the quality of text mining significantly. This aspect requires the attention of future researcher.

iii) Sentiment:- Sentiment and emotion analysis has gained significance prominence in the field of text mining due to the interest of governments and multinational companies to maintain a finger on the pulse of the public mood to win elections in the case of the former or just surprise their customers by the amount of insight about their preferences for the latter. Interestingly, market – prediction is very closely related to the mood of public or market participants as established by behavioral –economics. However, in case of the analysis of sentiment with regards to a product the anticipation of what apiece of text entails is far more straightforward than in the case of market prediction. There are no secretes as to whether a product review entails positive or negative emotions about it. However, even the best traders and investors can never be completely sure what market reaction to except as a result of a piece of news-text. Therefore, there is a lot of room for market –predictive sentiment investigation for future research.

iv) Text-mining component, textual-source or application market specialization:- In the future, the text –mining process should be broken down into its critical components like feature selections. Market – predictive text mining can also become even more specialized by focusing on a specific source of text. For ex. A specific social media outlet or news-source or text – role like news headlines. Moreover, there is a need for specialized research on each type of financial markets (stocks, bond, commodity, money, futures, derivatives, insurance, forex) or on each geographical location.

VI. CONCLUSION

Sentiment detection has a wide variety of applications in information systems, including classifying reviews, summarizing review and other real time applications.

The major systems for market prediction based on online text mining have been reviewed. The review was conducted on three major aspects. Namely:- preprocessing, machine learning and the evaluation mechanism. This work intended to accomplish:

i. Facilitation of integration of research activities from fields on the topic of market prediction based on online text mining.

ii. Provision of a study-framework to isolate the problem or different aspects of it.

iii. Submission of directional and theoretical suggestions for future research.

Advance in the field of market prediction can have the following implications

i. Investment banks and financial institutions as well as brokerage firms who are investing and trading in financial markets can use specialized market trend analysis and prediction systems that are developed using the insights gained in such targeted text mining research efforts as this work. The existence of such intelligent systems for those institutions assists with making better financial decisions and it is very important for returns on their investments and save losses.

ii. In today's global market, even more sophisticated insights into the financial markets is needed, because we have recently witnessed during the year 2008 financial crisis, can negatively impact on the millions of people around the world. Therefore it become imperative to pursue research in the field of market predictive text mining. That may bring about a much higher degree of confidence on comprehension of market-movements.

iii. The study opens up avenues for future research that could examine media effects on firm stock performance applying specific accounting or finance domain knowledge. Another area for future study is to explore the tone of the messages in various media sources and the extent of sentiment among the general public as compared to the sentiment among more sophisticated media practitioners as analysts.

REFERENCES

[1] C..Asvathiratham, The Influence Model: A Tractable representation for the Dynamics of networked Markov chains,Dept. of EECS,2000,MIT,Cambridge,2000,p188
[2] K.W.Cheung, J.T.Kwok, M.H.Law,K.C.Tsui, Mining customer product rating for personalized marketing, Decision support systems 35(2)(2003) 231-243
[3] J.M. Yang, X.Z. Hung,D. Zuang,,S.T.Zhang, The complex network analysis of competitive relationships between manufacturers in Foshan ceramic Industry cluster,2006 International Conference on management Science and Engineering,2006,pp. 2020-2023
[4] http// finance. Google,com

[5]http//finance.yahoo.com
[6]B.Pang, L.Lee, S.Vaithyanathan, Thumbs up sentiment classification using machine learning technique. Proceeding of the conference on empirical method in natural language processing(EMNLP),2002,PP 79-86
[7] http//zp.ioche.com
[8] http// www.youtube.com
[9]K.Ahmad, Y.Almas,Visualising sentiments in financial texts proceedings of the Ninth International conference on Information Visualisation(2005) 363-368
[10] P.Chaovalit, L.Zhou,Movie review mining: a comparison between supervised and unsupervised classification approaches ,Proceeding of the 38th Hawaii International conference on system sciences,2005
[11] P.D.Turney, Mining the web for synonyms: PMI-IR versus LSA on TOEFL, Procedding of the twelfth European conference on Machine Learning, Springer Verlag, Berlin,2001,pp. 491-502
[12] P.D.Turney, M.I.Littman,315-346,Measuring praise and criticism,inference of semantic orientation from association,ACM Transactions on information systems 21(2003) 315-346
[13] D. Xu S.Liao,Q.Li, Combining empirical experimentation and modeling techniques: a design research approach for personalized mobile advertising applications,Decision support systems 44(3)
(2008.)
[14] S.Yuan, A personalized and integrative comparision-shopping engine and its applications, Decision supports systems 34(2) (2003)
[15] V.Guralnik,G.Karypis, A. scalable algorithm for clustering protin sequence procedure workshop data mining in bioinformatics(BIOKDD),2001,pp 73-80
[16] K.F.Han, D.Baker, global properties of the mapping between local amino acid sequence and local structure in proteins,Proceedings of the natural Academy of Science of the united state of America(1996) 5814-5818
[17]W.Zhong,G.Altun, R.Harrison, P.C.Tai. Y.Pan ,Improved K-means clustering algorithm for exploring local protein sequence motifs representing comman structural property ,IEEE Transaction on NanoBioscience 4(3)(2005) 255-265
[18] M.Dash,H.Liu,Feature selection for classification, Intelligent data analysis 1(3) (1997) 131-156
[19]C.C.Freifeld, K.D.Mandl, ,B.Y.Reis, J.S.Brownstein, HealthMap: global infectious disease monitoring through automated classification and visualation of internet media reports,Journal of the American medical informatics Association 15(2008) 150-157
[20] T.Saegusa, T. Maruyama,Real-Time segmentation of color imageses based on the K-means CLUSTERING ON FPGA,International conference on field- programable.Technology, 2007,pp329-332.
[21] I.J.Khan,S.Shaikh,relationship algebra for computing in social network and social network based application 2006 IEEE WIC/ACM International Conference on web Intelligence 2006, pp 113-116
[22] N.Godbole,M.Srinivasaiah, S.Skiena, large scale SA for news and blogs ,Proceedings of the international conference on blogs and social media(ICWSM),2007.
[23] B.Pang.L.Lee, opinion mining and sentiment analysis,2008.
[24] B.Liu,Sentiment analysis and subjectivity, handbook of natural language processing(2010) 627-666.