Online Journal of Communication and Media Technologies
Volume: 7 – Issue: 1 January - 2017

Social Media Aided Sentiment Analysis in Forecasting

K.Nirmala Devi, Kongu Engineering College, India

Abstract

User generated contents on web and social media grow rapidly in this emerging information age. Social media provides a platform for people to create contents, share them and bookmark them in a tremendous way. The exponential growth of social media arouses much attention on the use of public opinion to make better decisions about a particular product or person or service. The social media like online forums, Twitter, Facebook, blogs and microblogs are proving to be extremely valuable resources for anticipating, detecting and forecasting significant societal events. It provides a lot of opportunities for users to voice their opinions openly. The analysis of sentiments obtained through social media along with wisdom of crowds can automatically compute the collective intelligence of future performance in many areas like stock trend forecasting, box office sales, hot topic detection, election outcomes and so on. The proposed research aims to perform forecasting based on user sentiments in social media regarding hotspots and stock forecasting.

Keywords: social media, forecasting, sentiment analysis, opinion mining, wisdom of crowds

Introduction

Social media provides a pool of collaborative knowledge, which also allows the understanding of collaborative pricing behavior on stock markets. In the era of information technology, various agents release the news and many people share their views and comments in the social media. The analysis of the sentiments obtained through social media along with wisdom of crowds can automatically compute the collective intelligence of future performance in many areas like stock trend forecasting (Nizer & Nivola 2011), box office sales, hot topic detection (Kuan-Yu et al 2007), predicting election outcomes and so on.

Social media has received a great attention by many users due to its easy access via portable devices for sharing a huge amount of data. Collaborative knowledge in social multimedia sites such as Flickr, Facebook, Youtube, Picasa, ImageShack and Photobucket are serving as an information platform for various social science studies because the image data have richer information that are useful for identifying the associated events from various sources (Xin et al 2010).The snippets of text in social media are a gold mine for companies and individuals who want to monitor their reputation and get timely feedback about their products and actions. Sentiment Analysis (SA) offers the organizations the ability to monitor different social media sites in real time and act accordingly. Marketing managers, campaign managers, politicians, equity investors and online shoppers are the direct beneficiaries of SA (Ronen 2013).

The tremendous growth of social media data has become a source of inspiration to many researchers in various disciplines like marketing and e-learning. The increasing popularity of social media encourages more and more users to participate in various online activities and produces data in an unprecedented rate. However, most of the data obtained from the social media are noisy, incomplete, highly unstructured and sometimes semi-structured in form and so it is more difficult to decipher them automatically. Hence, it is quite essential to understand and analyze these data for making a right decision in various forecasting application. The rest of this paper is organized as follows. The research works related to the proposed system is described in Section 2. The detailed proposed approach is discussed in Section 3.The experimental results and performance analysis are discussed in Section 4.The Section 5 contains the conclusion of the paper.

Related Works

Social media is the fastest growing phenomenon on the web, enabling millions of users to generate and share knowledge. With the paradigm shift in the usage of the Web 2.0 from information consumption to information production and sharing, numerous social media services have emerged. Online users can now conveniently express their opinions through news portals, forum discussions, reviews, messages, blogs and microblogs, thereby their willingness to engage in social interactions increase tremendously (Yanghui et al 2014). Findings on social media and understanding the contents discussed over social media have triggered researchers' interest on whether social media can be used as a signal for enhancing models in several tasks such as Human Computer Interaction (HCI), identifying political sentiment (Tumasjan el al 2011), predicting movie ratings and box office revenues, book sales, product recommendation and so on.There has been a considerable amount of research carried out in the hotspot detection based on various methods. Analysis of hotspots (Chen &Chundi 2008; Zhang & Wu 2009) of a topic is performed with various techniques such as classification, clustering (Andrew &Lawson 2010) and sample weighting.

Social Media and Sentiment Analysis

The extensive usage of social media has brought its pervasive impacts in different fields. This interesting phenomenon arises the present research to explore the impact of social media on a specific field. Nowadays, with a plenty of people on it the impact of social media on different aspects of the society has been more prominent. The social network sites and micro-blogging sites are considered to be a very good source of information because people share and discuss their opinions about a certain topic freely and besides popularity, social media has been playing an increasingly important role in predicting present or near future events.

In one of the early works to leverage social media for future prediction, Gruhl et al (2005) have explored the correlation of the mentioned rate of products in online chatter posts and their sales spikes. The analysis shows that the volume of blog postings can be used to predict spikes in actual consumer purchase decisions at the online retailer Amazon (Das &Chen 2007). However, the study does not analyze how sentiments in blog posts can impact the consumers' decision about a purchase.

Sentiment Analysis (SA)

A thought or view based on emotion is called a sentiment. SA or Opinion Mining (OM) is an automatic process of mining opinions, emotions and attitudes from text or speech by using Natural Language Processing (NLP). OM is an important sub discipline within data mining, NLP and web mining, which automatically extracts, classifies and understands the opinions generated by various users (Bing 2012).

Data mining is concerned with the process of automatically extracting novel and non-trivial information from unstructured text documents by combining techniques from text mining, Machine Learning (ML), NLP, Information Retrieval (IR) and knowledge management. The functionalities of data mining are classified into two categories according to the kinds of pattern to be found: descriptive and predictive. Descriptive mining characterizes the general properties of the data in the database, while predictive mining infer the current data in order to make predictions.Common data mining tasks involve document classification, summarization, clustering of similar documents, concept extraction and SA. Web mining is a sub discipline of data mining used to extract semi-structured data in the form of web content mining, web structure mining and web usage mining.SA techniques can basically be divided into ML based approach, lexicon based approach (Dang et al 2010) and hybrid approach. ML based approach uses linguistic features with famous algorithms like Support Vector Machine (SVM), Maximum Entropy (ME) and Naive Bayes (NB), while the lexicon based approaches the sentiment lexicon, which is a predefined sentiment term in the sentiment dictionary. Similarly, lexicon based approach is divided into dictionary based approach and corpus based approach. The dictionary based approach uses the opinion seed words with their synonyms and antonyms. The corpus based approach uses initial opinion seed word list, and then finds the additional opinion words in a huge corpus with context specific orientations. The hybrid approach combines both the ML and the lexicon based approaches.

SA techniques help to enhance the value of the existing information resources in many ways that can be useful in decision making, which is affected by the opinions formed by leaders and other people. During this decision making process, other people's thought has always been an important piece of information. When a person wants to buy a product online, he or she would typically start by searching for reviews and opinions written by the other people on various offerings. Similarly, gathering information on how people converse regarding

particular products can be helpful when designing marketing and advertising campaigns. The predictive sentiment analysis is an approach that uses SA to predict the changes in the interesting field based on opinions. The ability to predict the outcome of future events quickly and accurately is critical in today's business environment. For example, sales predictions and sensing of future consumer behaviors can help business people alter research and production planning. The users can face scientific challenges during the extraction of useful information from the vast data source due to the data diversity and lack of formal structure.

Hotspot Detection and Forecasting

A hotspot forum is defined as a forum that appears frequently over a time period and has bulk threads as well as posts. Nowadays, social media have revealed their predictive influence (Jasmina et al 2013) on a variety of domains, which motivates the use of its contents to identify hotspot forums (Nirmala Devi &MuraliBhaskaran 2012). More specifically, the popularity of the forum is determined by the burst nature of the threads, the comments and the views of the users. The hotness of a forum depends upon how often it covers hot terms and many discussions that contains those terms. Therefore, the hotspot forum detection and forecasting is one of the promising research areas in web mining.

Stock Forecasting

Stock markets are a major component of the world's economy because they provide a large platform for companies to raise more money easily. Forecasting the trends of a stock market is a very difficult and a highly complicated task because it is affected by many factors such as economic conditions, political events, investors' sentiments and so on. The stock market series are generally dynamic, non-parametric, noisy and chaotic by nature. The scientific challenges of extracting useful information from this vast source of data are great due to its diversity and lack of formal structure.

Frame Work of Opinion Mining System

OM framework includes various components like information discovery, preprocessing, transformation of data and feature extraction, feature selection, opinion processing and mining, interpretation and evaluation and opinion services. Figure 1.1 demonstrates the framework of the OM process.

Information Discovery

The OM system identifies the required data from a variety of information sources like Twitter, Facebook, forums and so on. Crawler technology is used to locate and get the required data from the data sources that are saved in the Opinion DataBase (ODB).

Preprocessing

Preprocessing is an extremely significant part in the OM system, which includes parsing, duplication elimination, tokenization and Part Of Speech (POS) tagging, Named Entity Recognition (NER) and so on. Exploiting the user generated repositories and mining useful information from them have become a highly challenging task in OM and knowledge discovery because of the heterogeneity nature of data. Hence it is difficult for the users to interpret. A preprocessing phase is necessary to improve the quality of data and to make feature extraction phase more reliable, wherein noise and irrelevant data are removed.

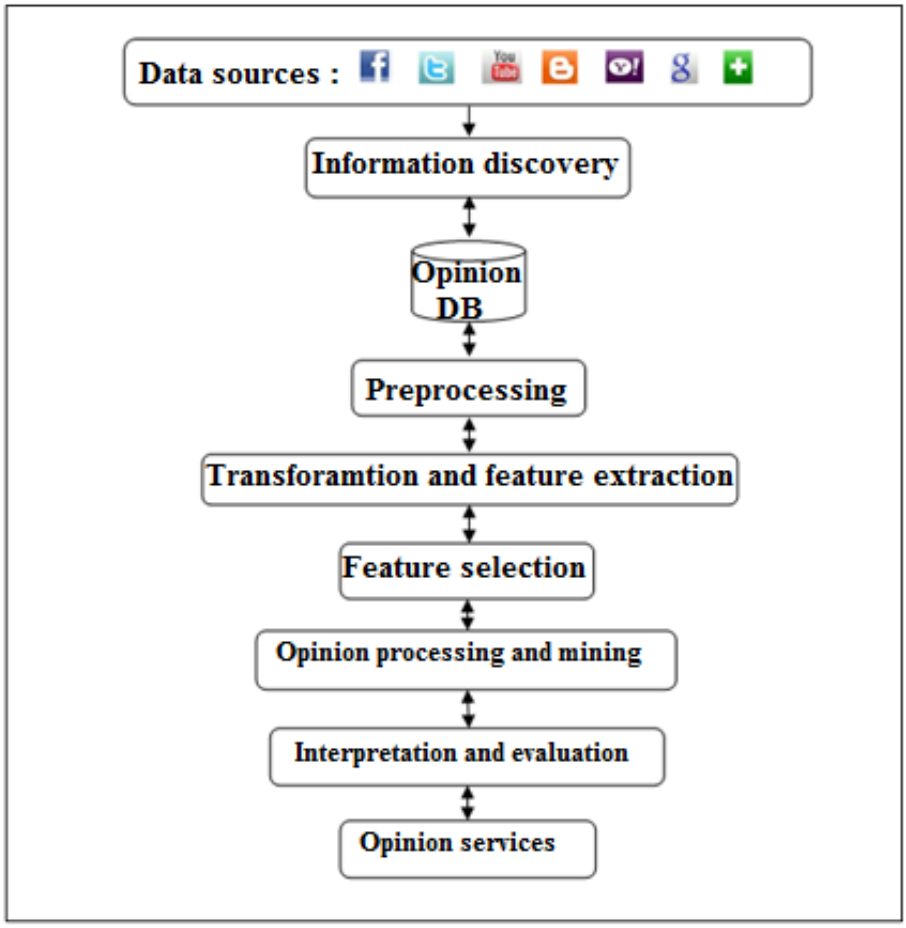

Figure 1. Framework of Opinion Mining System

Transformation of Data and Feature Extraction

In the transformation part, the data are transformed into appropriate form for mining by applying transformation techniques like normalization, smoothing and aggregation. A feature is defined as a function of one or more measurements and each one specifies some quantifiable properties of an object. Feature extraction is also a very important process used in clustering and classification process. Its purpose is to reduce the original data set by measuring certain properties.

Feature Selection

Feature selection is an essential component used in categorization and forecasting to identify the relevant features and to reduce the dimensionality of data to gain higher accuracy with low error rate (Feng et al 2012). After the extraction of features, a subset of most robust features has to be selected to improve the classification accuracy and to reduce the overall complexity. The main objective of a feature selection is obtain a feature space with low dimensionality, retention of sufficient information, enhancement of separability in feature space and comparability of features of the same category.

Opinion processing and mining

Opinion processing and mining is the most critical process in the entire framework of data mining task. There is a need to find the key and useful information from a huge amount of data like open forum data. Mining is an essential process wherein intelligent methods are applied to extract data patterns for interpretation and evaluation. Efficient statistical and ML techniques can be applied to process the enormous amount of online data. An emergent technique called emotional polarity computation also known as SA can be employed during online text mining. In opinion classification, the topic related words are not very important rather than opinion words that indicate positive or negative opinions, e.g. great, excellent, amazing, horrible, bad, worst and so on.

Interpretation and Evaluation of OM

Interpretation is the knowledge representation technique used to present the mined knowledge to the user while evaluation is used to identify the truly interesting patterns representing knowledge based on some interesting measures.

Opinion Services

Opinion services are used in many applications for prediction purpose like the following.

- Sentiment prediction
- Subjectivity detection
- Sentiment classification
- Product feature selection
- Aspect based sentiment summarization
- Text summarization for opinions
- Contrastive summarization

Opinion Spam Detection

Sentiment prediction is used to predict the polarity of the text; whether it is positive or negative. Subjectivity detection is a task of detecting whether the text is opinionated or not, while Sentiment Classification (SC) classifies the opinions into three categories such as 'positive', 'negative' and 'neutral'. Product feature selection is a task that extracts the product features from its review.

Aspect based sentiment summarization provides sentiment summary in the form of star ratings or scores of features. Text summarization generates a miniature form of sentences that summarize the reviews of any product. Contrast based viewpoint summarization puts an emphasis on contradicting opinions. Detecting opinion spam is concerned with identifying fake or bogus opinions from reviews.

Results and Discussion

The performance measures used for evaluating the proposed system are accuracy, precision and recall that are defined in Equations (1), (2) and (3).

$$Precision(P) = \frac{TP}{TP+FP} \quad (1)$$

$$Recall(R) = \frac{TN}{TP+FN} \quad (2)$$

$$Accuracy\ (A) = \left[\frac{No.of\ Right\ Predictions}{Total\ Predictions}\right] * 100 \quad (3)$$

Figure 2 presents the results of the accuracy measures of various models like Naïve Bayes (NB), J48, SVM and SVM-PSO for forecasting hotspot and stock, which includes the before feature selection as well as after feature selection. The proposed SVM-PSO technique has

proven to be able to generate better results when compared with the others (Hong and Xiaojun 2010; Jasmina et al 2013; Lee et al 2013; Nan and Desheng 2010) in the prediction of the hotspots as well as stock trends.

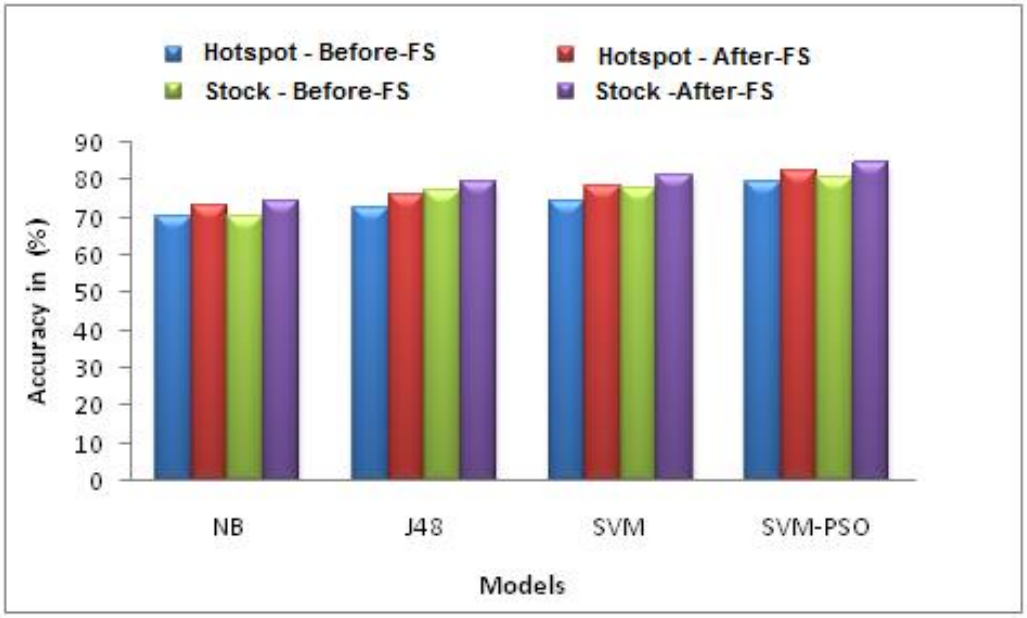

Figure 2. Results of Performance Measure based on Accuracy

Conclusion

In the present research, various methods have been developed for detecting the hotspot forums and for forecasting the stock trends. The results indicate that there is a correlation between social media sentiments and forecasting. The sentiment analysis along with wisdom of crowds can automatically compute the future performance of hotspot forums as well as stock trends. Thus, the efficient detection of hotspot forums and stock trend forecasting based on sentiment analysis with hybrid feature selection make social network members get benefited from effective decision making process. In future, the results generated from the hotspot detection can also be combined with market basket analysis to obtain comprehensive solution.

References

Andrew. B ,Lawson, 2010, 'Hotspot detection and clustering: ways and means', Environ Ecol Stat, vol. 17, pp.231–245.

Bing Liu 2012, 'Sentiment Analysis and Opinion Mining', Morgan & Claypool Publishers.

Chen Wei &Chundi Parvathi 2011, 'Extracting hot spots of topics from time-stamped documents', Data & Knowledge Engineering,vol. 70, 2011, pp. 642–660.

Dang, Y, Zhang, Y & Chen, H 2010, 'A Lexicon-Enhanced Method for Sentiment Classification: An Experiment on Online Product Reviews', IEEE Intelligent Systems, vol. 25, no. 4, pp. 46–53.

Das, SR, & Chen, MY 2007, 'Yahoo! for Amazon: Sentiment extraction from small talk on the web', Management Science, vol.53, pp.1375–1388.

Feng, G, Guo, J, Jing, BY, &Hao, L 2012, 'A Bayesian feature selection paradigm for text classification', Information Processing & Management, vol.48, pp.283–302.

Gruhl, D, Guha, R, Kumar, R, Novak, J & Tomkins, A 2005, 'The predictive power of online chatter', Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, pp. 78-87,New York.

Hong Liu &Xiaojun Li 2010,'Internet Public Opinion Hotspot Detection ResearchBased on K-means Algorithm', Proceedings of the ICSI , Part II, LNCS 6146, Springer-Verlag pp. 594–602.

JasminaSmailovic, Miha Grvcar1, Nada Lavrac& Martin Znidarsic 2013, 'Predictive Sentiment Analysis of Tweets: A Stock Market Application', HCI-KDD 2013, LNCS 7947, pp. 77–88, Springer-Verlag Berlin Heidelberg.

Kuan-Yu Chen, LuesakLuesukprasert, & Seng-cho T, Chou 2007, 'Hot Topic Extraction Based on Timeline Analysis and Multidimensional Sentence Modeling', IEEE Transactions on Knowledge and Data Engineering, vol.19, no. 8, pp.1016-1025.

Lee Zhong Zhen, Yun Huoy Choo, AzahKamilah Muda &Ajith Abraham 2013, 'Forecasting FTSE Bursa Malaysia KLCI Trend with Hybrid Particle Swarm Optimization and Support Vector Machine Technique', Proceedings of the World Congress on Nature and Biologically Inspired Computing, pp. 169-174.

Nirmala Devi, K &MuraliBhaskaran, V 2012, 'Sentiment Analysis for Online Forums Hotspot Detection', ICTACT Journal on Soft Computing, vol. 2, no. 2, pp.280-284.

Nizer, PSM &Nievola, JC 2012, 'Predicting published news effect in the Brazilian stock market', Expert Systems with Applications, vol.39, pp.10674–10680.

Ronen Feldman 2013, 'Techniques and Applications for Sentiment Analysis', Communications of the ACM, vol. 56, no. 4, pp. 82-89

Tumasjan, A, Sprenger, T, Sandner, P &Welpe, PI 2011, 'Election Forecasts With Twitter: How 140 Characters Reflect the Political Landscape', Social Science Computer Review, vol.29, no.4, pp. 402-418.

Xin Jin ,Andrew Gallagher, Liangliang Cao, Jiebo Luo & Jiawei Han 2010, 'The Wisdom of Social Multimedia: Using Flickr For Prediction and Forecast', Proceedings of the MM'10, Firenze, Italy.

Yanghui Rao , Qing Li , Xudong Mao & Liu Wenyin 2014, 'Sentiment topic models for social emotion mining', Information Sciences, vol. 266, pp. 90–100.