Available online at www.elixirpublishers.com (Elixir International Journal)

Computer Science and Engineering

Elixir Comp. Sci. & Engg. 56A (2013) 13751-13755

ISSN: 2229-712X

Do the sentiments on the net really matter?

Vanipriya[1] and Thammireddy.K[2]

[1]Department of Computer Science and Engineering, Sir M Visvesvarya Institute of Technology, Yelahanka, Bangalore, Karnataka, India

[2]Department of CSE, GIT, GITAM University, Vishakhapatnam, Andhra Pradesh, India.

ARTICLE INFO

Article history:
Received: 26 August 2012;
Received in revised form:
21 March 2013;
Accepted: 25 March 2013;

Keywords
Sentiment analysis,
Web mining,
Text extraction,
Text cleaning,
Naïve Bayesian Classifier,
Natural Language processing.

ABSTRACT

Our information-gathering behavior has always been to find out what other people think. Opinion-rich resources such as online review sites and personal blogs are growing rapidly which provide new opportunities and challenges, as people can actively use information technologies to seek out and understand the opinions of others. People share their experiences on-line, express their opinions, frustrations about anything. The huge amount of available data creates opportunities for automatic mining and analysis. It can be considered as a classification task: their feelings can be positive, negative or neutral. People don't directly express sentiment, They can also use a diverse range of other methods to express their emotions. Authors may mix objective and subjective information about a topic, or write down psots about other topics than the one we are investigating. And also there is a lot of noise in the data gathered from the Web pages. Due to this the task of automatic recognition of the sentiment in on-line text becomes more difficult. The content we are interested in this paper is, what is sentiment analysis, how the sentiment analysis was useful for a bank, how can it be applied in market intelligence and what are the challenges that could be faced in Indian scenario.

Introduction

Nowadays, there has been a rapid growth of web-content, especially on-line discussion groups, review sites and blogs. Some can be highly personal and typically express opinions. To organize this information, automatic text categorization and identification of sentiment polarity is very useful. Most work done in this field has been focused on topic based categorization, which is sorting the documents according to their subject of reference (SOR).

Sentiment classification is a special case of text categorization problem, where the classification is done on the basis of attitude expressed by the authors in discussion forums, blogs etc. Sentiment analysis requires a deep understanding of the document under analysis because the concern here is how the sentiment is being communicated.

Automatic sentiment analysis is a topic within information extraction that only recently received interest from the academic community. In the previous decade, a handful of articles have been published on this subject. It's only in the last five years that we've seen a small explosion of publications. The idea of automatic sentiment analysis is important for marketing research, where companies wish to find out what the world thinks of their product; for monitoring newsgroups and forums, where fast and automatic detection of flaming is necessary; for analysis of customer feedback; or as informative augmentation for search engines.

The automatic analysis of sentiments on data found on the Web is useful for any company or institution caring about quality control. For the moment, getting user feedback means bothering him or her with surveys on every aspect the company is interested in. The problems with this approach are making a survey for each product or feature; the format, distribution and timing of the survey (asking to send a form right after purchase might not be very informative); and the reliance on the goodwill of people to take the survey. This method can be made obsolete by gathering such information automatically from the World Wide Web, where the large amount of available data creates the opportunity to do so. One of the sources are blogs (short for "web logs"), a medium through which the blog owner makes commentaries about a certain subject or talks about his or her personal experiences, inviting readers to provide their own comments. Another source is the electronic discussion boards, where people can discuss all kinds of topics, or asks for other people's opinions. We define a topic as the subject matter of a conversation or discussion, e.g. an event in the media or a new model of car, towards which the writer can express his or her views.

There are several additional advantages to this approach. First, the people who share their views usually have more pronounced opinions than average, which are additionally influencing others reading them, leading to so called word-of-mouth marketing. Extracting these opinions is thus extra valuable. Second, opinions are extracted in real-time, allowing for quicker response times to market changes and for detailed time-based statistics that make it possible to plot trends over time.

Motivation

This paper is motivated by the malicious attack which was held against ICICI Bank. The attack stated that "Police probe 'smear campaign' against ICICI Bank". A 22-page complaint

Tele:
E-mail addresses: vani_hm72@yahoo.co.in

was submitted to police, alleging that rogue brokers spread malicious rumors about the bank's financial status.

Mumbai: Police in Mumbai said Tuesday they were investigating claims that a "bear cartel" of brokers tried to bring down ICICI Bank in a text message, email and Internet smear campaign. A 22-page complaint was submitted to police, alleging that rogue brokers spread malicious rumors about the bank's financial status, causing a run on some branches and sending share prices plummeting.

According to the complaint, on of the text messages read: "Kindly withdraw all your deposits and cash from your account in ICICI Bank as ICICI Bank already rushed to RBI for insolvency."

The bank likened the incident to a "new form of economic terrorism" designed to hit public confidence and compromise national economic interests, the Press Trust of India said.

With this private bank felt they were attacked with rumors. To get the support, the bank wanted to show to RBI and Govt. of India that the rumors are indeed, a malicious attack. So it had to substantiate its claims.

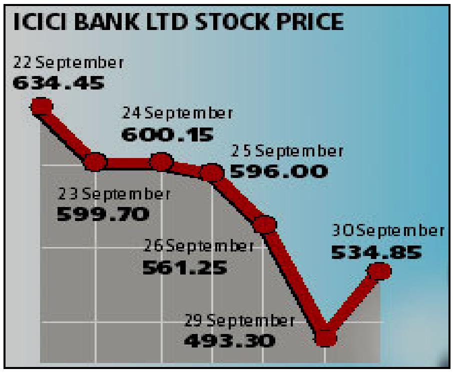
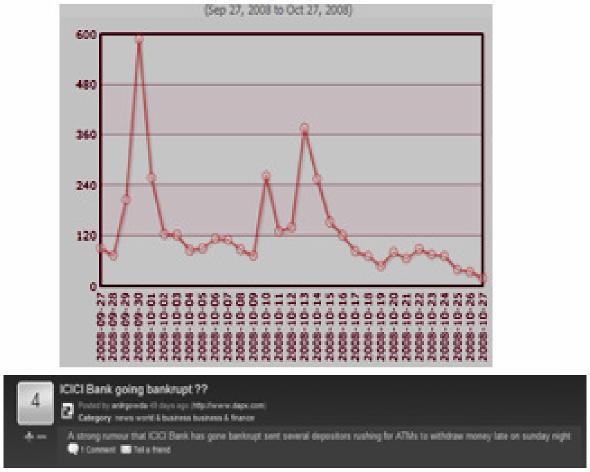


Fig 1: Stock price status of ICICI



ICICI Bank going bankrupt ??

Posted by anilgowda 43 days ago http://www.dapx.com

Category: news world & business business & finance

A strong rumour that ICICI Bank has gone bankrupt sent several depositors rushing for ATMs to withdraw money late on sunday night

1 Comment Tell a friend

Fig 2: The bank correlates the data on a time line

With the above problem, the following issues were recognized.

- The stock was plummeting; customers are withdrawing money in long queues.
- The bank has to provide evidence to the authorities for a speedy action.
- It noticed the rumors are spreading via SMS and online.
- Getting the SMS data is tricky. There are privacy issues and it cannot get until the cyber crime police is involved.
- The other option is to capture size-able online postings and use this as proof.

With the power of this information, the bank submits it all to the respective authorities. The bank responds to each of the posts.

Implementation details

The above problem finds its solution in the below two step procedure. They are:

a. Web mining

b. Sentiment Analysis with NLP

Web mining

Web mining is a technique to search, collate and analyze patterns in the data content of web sites by using traditional data mining techniques and attributes such as clustering and classification, association, and examination of sequential patterns.

The process of Web mining includes three important sub processes:

- Content mining
- Structure mining and
- Usage mining

Content mining is used to search, collate and examine data by search engine algorithms (this is done by using Web Robots).

Structure mining is used to examine the structure of a particular website and collate and analyze related data.

Usage mining is used to examine data related to the client end, such as the profiles of the visitors of the website, the browser used, the specific time and period that the site was being surfed, the specific areas of interests of the visitors to the website, and related data from the form data submitted during web transactions and feedback.

Sentiment Analysis

This is also known as opinion mining. It Attempts to identify the opinion/sentiment that a person may hold towards an object.

Several systems have been built which attempt to quantify opinion from product reviews. Pang, Lee and Vaithyanathan[10] perform sentiment analysis of movie reviews. Their results show that the machine learning techniques perform better than simple counting methods. They achieve an accuracy of polarity classification of roughly 83%. They identify which sentences in a review are of subjective character to improve sentiment analysis

Steps involved in Sentimental analysis

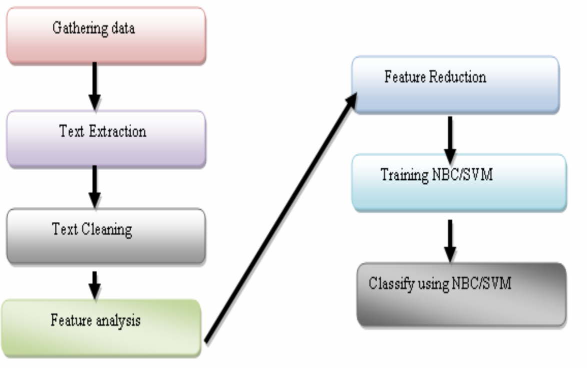


Fig 3: Steps involved in Sentimental analysis

Step 1: Gathering data from internet is solely based on the (SOR) Subject of Reference (e.g. ICICI bank). We use web mining techniques (ex. crawler) to gather all web pages where the SOR is mentioned.

Step 2: Text Extraction can be done in several data mining or text mining techniques starting from simple 'keyword matching' to 'DOM structure mining' to 'neural networks' methods. The major challenge here is that web documents are highly unstructured and no single method can give 100% clean text extraction for all documents.

Step 3: Text Cleaning is mostly heuristic based and case specific. By this we mean is to identify the unwanted portions in the extracted contents from Step 2 with respect to different kinds of web documents (e.g. News article, Blogs, Review, Micro Blogs etc) and then write simple cleanup codes based on that learning, which will remove such unwanted portions with high accuracy.

Step 4: Once we have the data corpus of clean documents from the previous steps, it is then put to the various knowledge processing engines. This data can be analyzed for feature analysis or business analytics or market research or consumer buzz trends or consumer sentiment analysis etc. depending on the needs. Various techniques are used for each such purpose. For example, Inverse Document Frequency (TF-IDF) technique can be used to gather a pool of various features for the SOR, which in case of ICICI bank can get a pool as (customer service, credit card, recovery agent, customer satisfaction etc). This is called feature analysis. From this pool one can determine things like how many consumers talk about recovery agent while talking about credit card compared to how many consumers talk about customer satisfaction while talking about credit cards.

Step 4.1: Obviously the above technique for Feature analysis will throw up few unwanted features in the pool which needs to be removed from final analysis. This can be done by feature mapping from pool with keywords representing the various well-known features. These features can also be pre-defined by the SOR.

Step 5: Sentiment Analysis of web documents can be defined as the consumer opinion expressed through online medium e.g. Blog or review. Now days a consumer can choose to post his/her sentiment about a particular brand/product/feature online which can be categorized broadly as Positive, Negative or Neutral. The web documents where such sentiment has been expressed can be referenced for various analytical / actionable causes by that brand representative. For example, considerable amount of negative sentiment expressed by consumers about customer service of ICICI bank can be actionable insight for ICICI bank, in which case ICICI might want to restructure its customer service to give better customer satisfaction and thus tend to reduce negative sentiment about it in the net. Sentiment analysis can be done on the clean extracted web documents in two manners - manual rating or automated rating of such web documents. While manual rating is a near perfect method to do it, but it is a slow process when the volume of web documents is too high. Whereas automated system will be much faster method, but is bound to lack accuracy since it is effectively machine learning and deriving human sentiments through user generated content. Also, language barrier is a major challenge for automated sentiment analysis. Nevertheless, extensive research work on Natural Language Processing has addressed such challenges well and reasonably high performance machine learning techniques have evolved which can do sentiment analysis of web documents. The two most efficient such techniques are Naive Bayesian Classifier (NBC) and Support Vector Machines (SVM). These machine learning algorithms requires a learning corpus to first train on then from that training it can derive sentiment from web documents.

Step 5.1: Training NBC/SVM is fairly straightforward and well studied technique, where first a manually rated corpus of several thousands of web documents is generated, categorizing them as either negative, positive or neutral for a given SOR. This corpus is then fed to the NBC/SVM engines which generate several measuring parameters for a given document to fall into one of the three categories [negative, positive or neutral].

Step 5.2: Once the NBC/SVM engines are trained, they can now be used to categorize rest of the web documents using the parameters generated by them . The accuracy definitely won't be 100% but several layers of training and tuning can increase and optimize this accuracy.

NBC for classification (Naive Bayesian classifier)

A naive Bayes classifier is a term in Bayesian statistics dealing with a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model".

In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 4" in diameter. Even though these features depend on the existence of the other features, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple.

Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without believing in Bayesian probability or using any Bayesian methods.

In spite of their naive design and apparently over-simplified assumptions, naive Bayes classifiers often work much better in many complex real-world situations than one might expect. Recently, careful analysis of the Bayesian classification problem has shown that there are some theoretical reasons for the apparently unreasonable efficacy of naive Bayes classifiers (Zhang04). An advantage of the naive Bayes classifier is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

Challenges in sentiment classification

Given the multitude of potential applications, researchers have been devoting more and more attention to sentiment analysis. Much of the current work is devoted to *classification* problems: determining whether a particular document or portion thereof is subjective or not, and/or determining whether the opinion it expresses is positive or negative. At first blush, this might not appear so hard: one might expect that we need simply look for obvious sentiment-indicating words, such as "great". The difficulty lies in the richness of human language use. First, there can be an amazingly large number of ways to say the same thing (especially, it seems, when that thing is a negative perception); this complicates the task of finding a high-coverage set of indicators. Furthermore, the same indicator may admit several different interpretations. Consider, for example, the following sentences:

- This laptop is a great deal.

- A great deal of media attention surrounded the release of the new laptop model.
- If you think this laptop is a great deal, I've got a nice bridge you might be interested in.

Each of these sentences contains the three words "a great deal", but the opinions expressed are, respectively, positive, neutral, and negative. The first two sentences use the same phrase to mean different things. The last sentence involves sarcasm, which, along with related rhetorical devices, is an intrinsic feature of texts from unrestricted domains such as blogs and newsgroup postings.

In general, researchers have adopted one of two approaches to meeting the challenges that sentiment analysis presents. Many groups are working to directly improve the selection and interpretation of indicators through the incorporation of linguistic knowledge; given the subtleties of natural language, such efforts will be critical to building operational systems. Others have been pursuing a different tack: employing *learning algorithms* that can automatically infer from text samples what indicators are useful. Besides being potentially more cost-effective, more easily ported to other domains and languages, and more robust to grammatical mistakes, learning-based systems can also discover indicators that humans might neglect. For example, in our own work, we found that the phrase "still," (comma included) is a better indicator of positive sentiment than "good" --- a typical instance of use would be a sentence like "Still, despite these flaws, I'd go with this laptop". Nevertheless, it bears repeating that incorporating deep knowledge about language will be absolutely crucial to developing systems capable of high-quality (as opposed to merely high-throughput) sentiment analysis. Both the linguistic and the learning approach have considerable merits; it seems very safe to say that the community will need to turn towards finding ways to combine their advantages.

Especially in India most of the people use 'HINGLISH' form of language for an eg.'ICICI is big chor '.In this case sentiment analysis will become more complicated.

Related problems, New directions

The classification problems discussed above only involve the determination of sentiment. However, there is growing interest in capturing interactions between *subjectivity* and *subject* --- we not only need to know what an author's opinion is, but what that opinion is about. For example, while in a broad sense a review of a particular laptop is only about one topic (the laptop itself), it almost surely discusses various specific aspects of the machine. We would ideally like a sentiment-analysis system to reveal whether there are particular features that the review's author disapproves of even if his or her overall impression was positive.

Another interesting research direction of potentially great importance is to integrate into sentiment analysis the notion of the *status* of an opinion holder, perhaps via adaptation of the hubs-and-authorities techniques used in Web search or link-analysis methods in reputation systems. For example, we might want to identify *bellwethers* --- thought leaders with enough influence that others explicitly adopt their opinions --- or *barometers* --- those whose opinions are generally held by the majority of the population of interest. Tracking the views of these two types of people could both streamline and enhance the process of gathering business intelligence to a large degree. Surely that sounds like a great deal!

Applications

- Businesses and organizations:
  - product and service bench marking.
  - market intelligence.
  - Business spends a huge amount of money to find consumer sentiments and opinions.
  - Consultants, surveys and focused groups, etc
- Individuals: interested in other's opinions when
  - purchasing a product or using a service,
  - finding opinions on political topics
- Ads placements: Placing ads in the user-generated content
  - Place an ad when one praises a product.
  - Place an ad from a competitor if one criticizes a product.
- Opinion retrieval/search: providing general search for opinions
- Product review mining: What features of the ThinkPad T43 do customers like and which do they dislike?
- Review classification: Is a review positive or negative toward the movie?
- Tracking sentiments toward topics over time: Is anger ratcheting up or cooling down?
- Prediction (election outcomes, market trends): Will Clinton or Obama win?
- Expressive text-to-speech synthesis
- Text semantic analysis Text summarization

Conclusions

The World Wide Web is today the major source of dataand information for all domains. It is not only anaccessible and searchable information source but alsoone of the most important communication channels, almost a virtual society. Web mining is an important and challenging activity that aims to discover new, relevant and reliable information and knowledge by investigating the web structure, its content and its usage. In our paper two main AI techniques: the multi-agent systems and swarm intelligence, with some of their applications in web mining. The mining tasks are so complex that they cannot be efficiently performed without the support of appropriate advanced AI techniques.

References:

1.Pang, Bo, Lillian Lee and Shivakumar Vaithyanathan. 2002.Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*

2.Hu, M., and Liu, B. 2004. Mining and summarizing customer reviews. In *Proceedings of KDD '04, the ACM SIGKDD international conference on Knowledge discovery and data mining*. Seattle, US: ACM Press.

3. J. Reilly and L. Seibert, "Language and Emotion," Handbook of Affective Science, R.J. Davidson, K.R. Scherer, and H.H. Goldsmith, eds., Oxford Univ. Press, 2003.

4. A. Neviarouskaya, H. Prendinger, and M. Ishizuka, "Textual Affect Sensing for Sociable and Expressive Online Communication," Proc. Second Int'l Conf. Affective Computing and Intelligent Interaction, 2007.

5. S. Argamon, K. Bloom, A. Esuli, and F. Sebastiani, "Automatically Determining Attitude Type and Force for Sentiment Analysis,"
Proc. Third Language and Technology Conf., 2007

6. P.D. Turney and M.L. Littman, "Measuring Praise and Criticism:Inference of Semantic Orientation from Association," ACM Trans. Information Systems, vol. 21, no. 4,, 2003.

7. S.-M. Kim and E. Hovy, "Determining the Sentiment of Opinions," Proc. Conf. Computational Linguistics,2004.

8. R. Agrawal, S. Rajagopalan, R. Srikant, and Y. Xu, "Mining newsgroups using networks arising from social behavior," in *Proceedings of WWW*, 2003.

9. N. Archak, A. Ghose, and P. Ipeirotis, "Show me the money! Deriving the pricing power of product features by mining consumer reviews," in *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2007.

10. M. Bautin, L. Vijayarenu, and S. Skiena, "International sentiment analysis for news and blogs," in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, 2008.

11. J. Berger, A. T. Sorensen, and S. J. Rasmussen, "Negative publicity :When is negative a positive?," Manuscript. PDF file's last modification date: October 16, 2007, *URL*: http://www.stanford.edu/□asorense/papers/ Negative Publicity.pdf, 2007.

12. Y. Chen and J. Xie, "Online consumer review: Word-of-mouth as a new element of marketing communication mix," *Management Science*, vol. 54, 2008.

13. comScore/the Kelsey group, "Online consumer-generated reviews have significant impact on offline purchase behavior," Press Release, http://www.comscore.com/press/release.asp?press=1928, November 2007.

14. J. G. Conrad and F. Schilder, "Opinion mining in legal blogs," in *Proceedings of the International Conference on Artificial Intelligence and Law (ICAIL)*,New York, NY, USA: ACM, 2007.

15. C. Dellarocas, "The digitization of word-of-mouth: Promise and challenges of online reputation systems," *Management Science*, vol. 49,2003. (Special issue on e-business and management science).

16. L. Dini and G. Mazzini, "Opinion classification through information extraction," in *Proceedings of the Conference on Data Mining Methods and Databases for Engineering, Finance and Other Fields (Data Mining)*, 2002.

17. W. Duan, B. Gu, and A. B. Whinston, "Do online reviews matter? — An empirical investigation of panel data," Social Science Research Network (SSRN) Working Paper Series, http://ssrn.com/paper=616262, version as of January, 2005.

18. C. Forman, A. Ghose, and B. Wiesenfeld, "Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets," *Information Systems Research*, vol. 19, 2008.

19. T. Fukuhara, H. Nakagawa, and T. Nishida, "Understanding sentiment of people from news articles: Temporal sentiment analysis of social events," in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, 2007.

20. M. Gamon, A. Aue, S. Corston-Oliver, and E. Ringger, "Pulse: Mining customer opinions from free text," in *Proceedings of the International Symposium on Intelligent Data Analysis (IDA),* number 3646 in *Lecture Notes in Computer Science.* 2005.

21. R. Ghani, K. Probst, Y. Liu, M. Krema, and A. Fano, "Text mining for product attribute extraction," *SIGKDD Explorations Newsletter*, vol. 8, 2006.

22. N. Godbole, M. Srinivasaiah, and S. Skiena, "Large-scale sentiment analysis for news and blogs," in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, 2007.

23. M. L. Gregory, N. Chinchor, P. Whitney, R. Carter, E. Hetzler, and A. Turner,"User-directed sentiment analysis: Visualizing the affective content of documents," in *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, Sydney, Australia, July 2006.

24. R. V. Guha, R. Kumar, P. Raghavan, and A. Tomkins, "Propagation of trust and distrust," in *Proceedings of WWW*, 2004.

25. Hankin, "The effects of user reviews on online purchasing behavior across multiple product categories," Master's final project report,UC Berkeley School of Information, http://www.ischool.berkeley.edu/files/lhankin report.pdf, May 2007.

26. N. Godbole, M. Srinivasaiah, and S. Skiena, "Large-scale sentiment analysis for news and blogs," in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, 2007.

27. T. Hoffman, "Online reputation management is hot — but is it ethical?" Computerworld, February 2008.

28. D. Hopkins and G. King, "Extracting systematic social science meaning from text,". Manuscript available at http://gking.harvard.edu/files/words.pdf,2007

29. M. Hu and B. Liu, "Mining opinion features in customer reviews," in *Proceedings of AAAI*, 2004.

30G. Jin and A. Kato, "Price, quality and reputation: Evidence from an online field experiment," *The RAND Journal of Economics*, vol. 37, 2006.

31. S.-M. Kim and E. Hovy, "Determining the sentiment of opinions," in *Proceedings of the International Conference on Computational Linguistics (COLING)*, 2004.

32. S.-M. Kim and E. Hovy, "Automatic identification of pro and con reasons inonline reviews," in *Proceedings of the COLING/ACL Main Conference Poster Sessions*, 2006.

33. L.-W. Ku, Y.-T. Liang, and H.-H. Chen, "Opinion extraction, summarization and tracking in news and blog corpora," in *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, 2006.

34. S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima, "Mining product reputations on the Web," in *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2002. (Industry track).

35. D.-H. Park, J. Lee, and I. Han, "The effect of on-line consumer reviews on consumer purchasing intention: The moderating role of involvement," *International Journal of Electronic Commerce*, vol. 11, (ISSN 1086-4415), 2007.

36. *n.wikipedia.org/wiki/Web_mining*

37. S. Chakrabarti. *Data mining for hypertext*: A tutorial survey. ACM SIGKDD Explorations, 1(2):1 11, 2000

37."Comparative Experiments on Sentiment Classification for Online Product Reviews", Hang Cui,