Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

*IJCSMC, Vol. 2, Issue. 7, July 2013, pg.231 – 239*

RESEARCH ARTICLE

# Sentiment Analysis Based Approaches for Understanding User Context in Web Content

**M. SAKTHIVEL[1], G. HEMA[2]**

[1]Department of Computer Science, Periyar University, Tamil Nadu, India

[2]Department of Computer Science, Periyar University, Tamil Nadu, India

*[1] msakthivelpgp@gmail.com; [2] johnmary.hema@gmail.com*

***Abstract— In our day to day lives, we highly value the opinions of friends in making decisions about issues like which brand to buy or which movie to watch. With the increasing popularity of blogs, online reviews and social networking sites, the current trend is to look up reviews, expert opinions and discussions on the Web, so that one can make an informed decision. Sentiment analysis, also known as opinion mining is the computational study of opinions, sentiments and emotions expressed in natural language for the purpose of decision making. Sentiment analysis applies natural language processing techniques and computational linguistics to extract information about sentiments expressed by authors and readers about a particular subject, thus helping users in making sense of huge volume of unstructured Web data. Applications like review classification, product review mining and trend prediction benefit from sentiment analysis based techniques. This paper presents a study of different approaches in this field, the state of the art techniques and current research in Sentiment Analysis based approaches for understanding user's context.***

***We show that information about social relationships can be used to improve user-level sentiment analysis. The main motivation behind our approach is that users that are somehow "connected" may be more likely to hold similar opinions; therefore, relationship information can complement what we can extract about a user's viewpoints from their utterances. Employing Twitter as a source for our experimental data, and working within a semi-supervised framework, we propose models that are induced either from the Twitter follower/follower network or from the network in Twitter formed by users referring to each other using "@" mentions. Our transductive learning results reveal that incorporating social-network information can indeed lead to statistically significant sentiment classification improvements over the performance of an approach based on Support Vector Machines having access only to textual features.***

***Key Terms: - opinion mining; computational linguistics to extract information; semi-supervised framework***

## I. INTRODUCTION

Data mining generally refers to a method used to analyze data from a target source and compose that feedback into useful information. This information typically is used to help an organization cut costs in a particular area, increase revenue, or both. Often facilitated by a data-mining application, its primary objective is to identify and extract patterns contained in a given data set.

***Kinds of information collecting***

- ***Business transactions****:* Every transaction in the business industry is (often) "memorized" for perpetuity. Such transactions are usually time related and can be inter-business deals such as purchases, exchanges, banking, stock, etc., or intra-business operations such as management of in-house wares and assets.
- ***Scientific data****:* Whether in a Swiss nuclear accelerator laboratory counting particles, in the Canadian forest studying readings from a grizzly bear radio collar, on a South Pole iceberg gathering data about

oceanic activity, or in an American university investigating human psychology, our society is amassing colossal amounts of scientific data that need to be analyzed. Unfortunately, we can capture and store more new data faster than we can analyze the old data already accumulated.

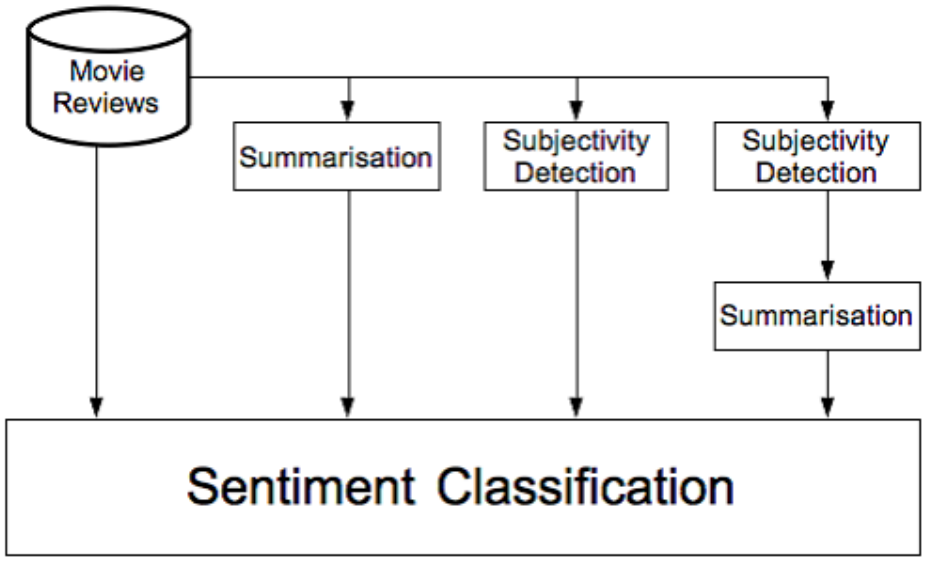
- ***Medical and personal data****:* From government census to personnel and customer files, very large collections of information are continuously gathered about individuals and groups. **Surveillance video and pictures**: With the amazing collapse of video camera prices, video cameras are becoming ubiquitous. Video tapes from surveillance cameras are usually recycled and thus the content is lost. However, there is a tendency today to store the tapes and even digitize them for future use and analysis.
- ***Satellite sensing****:* There is a countless number of satellites around the globe: some are geo-stationary above a region, and some are orbiting around the Earth, but all are sending a non-stop stream of data to the surface. NASA, which controls a large number of satellites, receives more data every second than what all NASA researchers and engineers can cope with. Many satellite pictures and data are made public as soon as they are received in the hopes that other researchers can analyze them.
- ***Games****:* Our society is collecting a tremendous amount of data and statistics about games, players and athletes. From hockey scores, basketball passes and car-racing lapses, to swimming times, boxers pushes and chess positions, all the data are stored. Commentators and journalists are using this information for reporting, but trainers and athletes would want to exploit this data to improve performance and better understand opponents.
- ***Digital media****:* The proliferation of cheap scanners, desktop video cameras and digital cameras is one of the causes of the explosion in digital media repositories.
- ***CAD and Software engineering data****:* There are a multitude of Computer Assisted Design (CAD) systems for architects to design buildings or engineers to conceive system components or circuits. These systems are generating a tremendous amount of data. Moreover, software engineering is a source of considerable similar data with code, function libraries, objects, etc., which need powerful tools for management and maintenance.
- ***Virtual Worlds****:* There are many applications making use of three-dimensional virtual spaces. These spaces and the objects they contain are described with special languages such as VRML.
- ***Text reports and memos (e-mail messages)****:* Most of the communications within and between companies or research organizations or even private people, are based on reports and memos in textual forms often exchanged by e-mail. These messages are regularly stored in digital form for future use and reference creating formidable digital libraries.

## II. A SURVEY OF EXISTING APPROACHES

Current research offers several interesting approaches to the challenge posed by SA. Abstracting away from specific implementations, these approaches can be classified into 3 categories: the lexical-phrasal approach, the compositional semantics approach, and the Machine Learning approach. Clearly, these approaches are not mutually exclusive. Indeed, systems employing all three have been implemented to varying degrees of success.

The strategies by which these approaches are implemented may vary significantly from system to system. For example, feature classification in Machine Learning based systems can be based on the authors' self-assessment (e.g. starred reviews as in Turney 2002; Pang 3 et al 2002; Finn & Kushmerick 2003; Kushal et al 2003) corpora tagged by human judges (e.g. Multi-Perspective Question Answering (MPQA) corpus; Wiebe et-al, 2001; Choi & Cadie, 2008) and even on exogenous sources (e.g. in Koppel & Shtrimberg, 2009, the prices of stocks serve as the judge). Likewise, the compilation of relevant lexicons and the definition of compositional semantic rules are subject to significant variation.




**Sentiment Analysis Steps**

Suppose we are interested in deriving the sentiment or opinion of various digital cameras across dimensions such as price, usability, and features.

**Step 1: Fetch, Crawl, and Cleanse**

Comments about digital cameras might be available on gadget review sites or in discussion forums about digital cameras, as well as in specialized blogs. Data from all of these sources needs to be collected to give a holistic view of all the ongoing discussions about digital cameras. Web crawlers—simple applications that grab the content of a Web page and store it on a local disk—fetch data from the targeted sites. The downloaded Web pages are in HTML format, so they need to be cleansed to retain only the textual content and the remaining HTML tags used for rendering the page on the Web site.

**Step 2: Text Classification**

The sites from which data is fetched might contain extra information and discussions about other electronic gadgets, but our current interest is limited to digital cameras. A text classifier determines whether the page or discussions on it are related to digital cameras; based on the decision of the classifier, the page is either retained for further analysis or discarded from the system.

**Step 3: Entity Extraction**

Entity extraction involves extracting the entities from the articles or discussions. In this example, the most important entity is the name or model of the digital camera—if the name is incorrectly extracted, the entire sentiment or opinion analysis becomes irrelevant.

There are three major approaches for entity extraction:

**Dictionary or taxonomy:** A dictionary or taxonomy of available and known models of digital cameras is provided to the system. Whenever the system finds a name in the article, it tags it as a digital camera entity. This technique, though simple to set up, needs frequent updates on every subsequent model launch, so it's not robust.

**Rules:** A digital camera model name has a certain pattern, such as Canon Power Shot A540. Therefore, a rule may be written to tag any alphanumeric token following the string "Canon Power Shot" as a digital camera model. Such techniques are more robust than the dictionary-based method, but if Canon decides to launch a new model, say the SuperShot, such rules must be updated manually.

***Machine learning:*** This algorithm learns the extraction rules automatically based on a sample of articles with the entities properly tagged. The rules are learned by forming graphical and probabilistic models of the entities and the arrangement of other terms adjoining them. Popular machine learning models for entity extraction are based on hidden Markov models (HMM) and conditional random fields (CRF).

***Step 4: Sentiment Extraction***

Sentiment extraction involves spotting sentiment words within a particular sentence. This is typically achieved using a dictionary of sentiment terms and their their semantic orientations. There are obvious limitations to the dictionary-based approach. For example, the sentiment word "high" in the context of "price" might have a negative polarity, whereas "high" in the context of "camera resolution" will be of positive polarity.

Once an entity of interest (for example, the digital camera model or sentiment word) is identified, structured sentiment is extracted from the sentence in the form of {model name, score}, where score is the positive or negative polarity value of the identified sentiment word in the sentence. If some dimension (such as "price" or "resolution") is also found in the sentence, then the sentiment is extracted in the form of {model name, dimension, score}. We may also choose to report the source name or source ID to associate the extracted sentiment back to that source.

***Step 5: Sentiment Summary***

The raw sentiments extracted in Step 4 come from individual sentences that are specific to certain entities. To make the data meaningful for reporting, it must be aggregated. One of the obvious aggregations in the context of digital cameras will be model-name-based aggregation—in this case, all of the positive, negative, or neutral entries in the database are grouped together. Again, model- and dimension-based sentiment aggregation would allow the discovery of detailed, dimension-wise sentiment distribution for every model. Based on the reporting needs, different levels of aggregation and summarization need to be carried out and stored in a database or data warehouse.

***Step 6: Reports/Charts***

Reports and charts can be generated directly from the database or data warehouse where the aggregated data is stored in a structured format. Such reporting falls under the purview of traditional BI and reporting, and is not related to the core sentiment analysis steps.

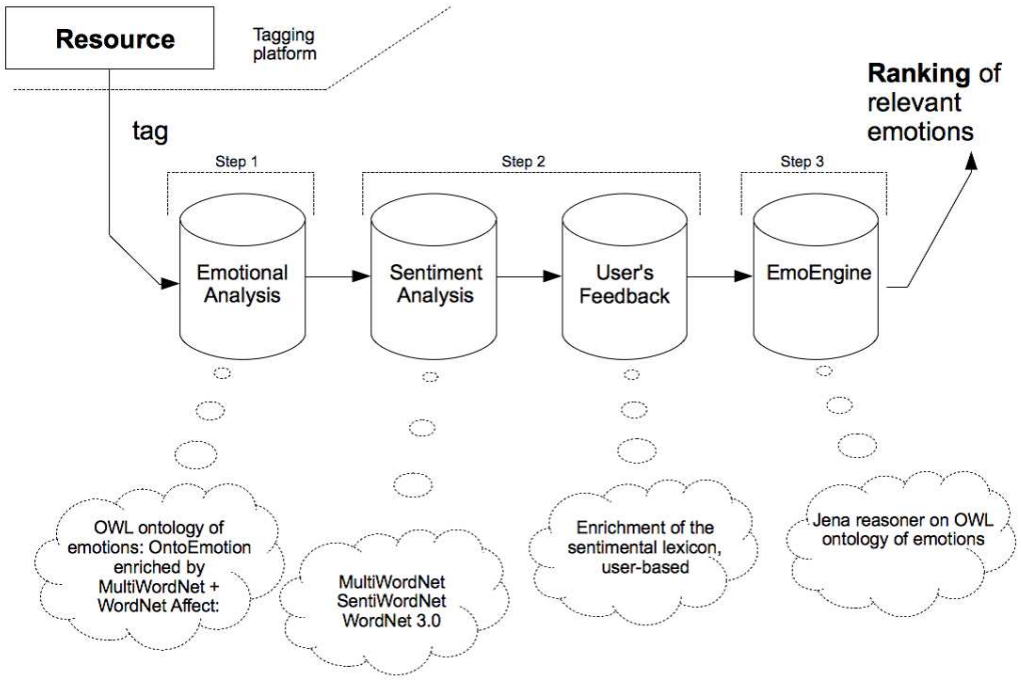
The steps described above have been used to transform the unstructured textual data in blogs and forums to structured, quantifiable numeric sentiment data related to the entity of interest.

## III. ARSEMOTICA

This section describes the architecture of ArsEmotica, the application software that we developed. The analysis steps that we are about to describe rely on a preprocessing phase in which tags are filtered so as to eliminate flaws like spelling mistakes, badly accented characters, and so forth. Figure 1 reports the three main Fig. 1 ArsEmotica overall architecture. steps that characterize the computation after the pre-processing:

***Step 1:*** Checking tags against the ontology of emotions. This step checks whether a tag belongs to the ontology of emotions. Tags belonging to the ontology are immediately classified as "emotional".




***Step 2:*** Checking tags with SentiWordNet. Tags that do not correspond to terms in the ontology are further analyzed by means of SentiWordNet, in order to distinguish objective tags, which do not bear an emotional meaning, from subjective and, therefore, affective tags. The latter will be the only ones presented to the user in order to get a feedback on which emotional concept they deliver.

***Step 3:*** Ranking of Emotions. Based on data collected in the previous steps, the tool ranks the emotions associated by the users to the resource. The following sections explain in details how the extraction of an emotional semantics is performed.

***Steps that characterize the computation after the pre-processing:***

***Step 1:*** Checking tags against the ontology of emotions. This step checks whether a tag belongs to the ontology of emotions. Tags belonging to the ontology are immediately classified as "emotional".

***Step 2:*** Checking tags with SentiWordNet. Tags that do not correspond to terms in the ontology are further analyzed by means of SentiWordNet, in order to distinguish objective tags, which do not bear an emotional meaning, from subjective and, therefore, affective tags. The latter will be the only ones presented to the user in order to get a feedback on which emotional concept they deliver.

***Step 3:*** Ranking of Emotions. Based on data collected in the previous steps, the tool ranks the emotions associated by the users to the resource.

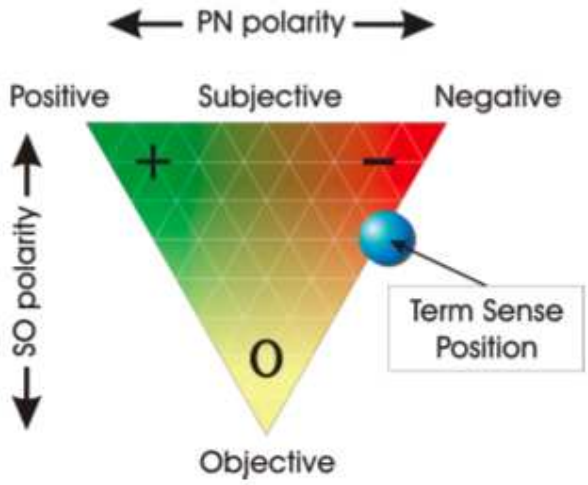
## IV. METHODOLOGIES

A wide range of tools and techniques are used to tackle the goals described above. This section describes some of the most common and interesting ones. First, Machine Learning and Part-Of-Speech tagging will be discussed, since these are very powerful tools that are most often used in Sentiment Analysis. Then specific techniques and approaches for tackling each of the tasks described in the previous section will be addressed.

### *A. Classification*

Many of the tasks in Sentiment Analysis can be thought of as classification.




**a. Term Presence Vs Frequency**

Traditional Information Retrieval systems have long emphasized the importance of term frequency. The famous TF-IDF (Term Frequency - Inverse Document Frequency) measure is well-used in modeling documents. The intuition is that terms that often appear in the document but seldom in the whole collection are more informative as to what the document is about as compared to the terms mentioned just once.

**b. n-grams**

Term positions are also important in document representation for Sentiment Analysis. The position of terms determines, and sometimes reverses, the polarity of the phrase.

**c. Part-of-Speech**

As mentioned earlier, it has been determined that adjectives are good indicatives of sentiment in text (Hatzivassiloglou and Wiebe, 2000; Benamara et al., 2007), and in the past decade they have been commonly exploited in Sentiment Analysis.

**d. Syntax**

Syntax information has also been used in feature sets, though there is still discussion about the merit of this information in Sentiment classification (Pang and Lee, 2008).

**e. Negations**

Negations have been long known to be integral in Sentiment Analysis. The usual bag-of-words representation of text disconnects all of the words, and considers sentences like “I like this book” and “I don’t like this book” very similar, since only one word distinguishes one from the other. But when talking about sentiment, a negation flips the polarity of a whole phrase.

***B. Identifying the semantic orientation of words***

One of the most basic tasks in Sentiment Analysis is identifying the semantic orientation (the polarity and objectivity) of a word. A variety of techniques have been used, which can be roughly categorized in the following: _ using a lexicon, constructed manually or automatically _ using some statistical techniques such as looking at concurrence of a word with a word of a known polarity _ using training documents, labeled or unlabeled, as a source of knowledge about the polarity of terms within the collection.

## V. RESULTS

***Experiments***

We conduct experiments on a movie review dataset collected from the IMDb archive. We use an n-gram model to represent the total review and its subjective excerpt as a feature vector to the classifier. Each n-gram is weighted using the tfidf score. We predict the overall polarity of a review as positive or negative.

***A. Experimental setup***

***a. Datasets***

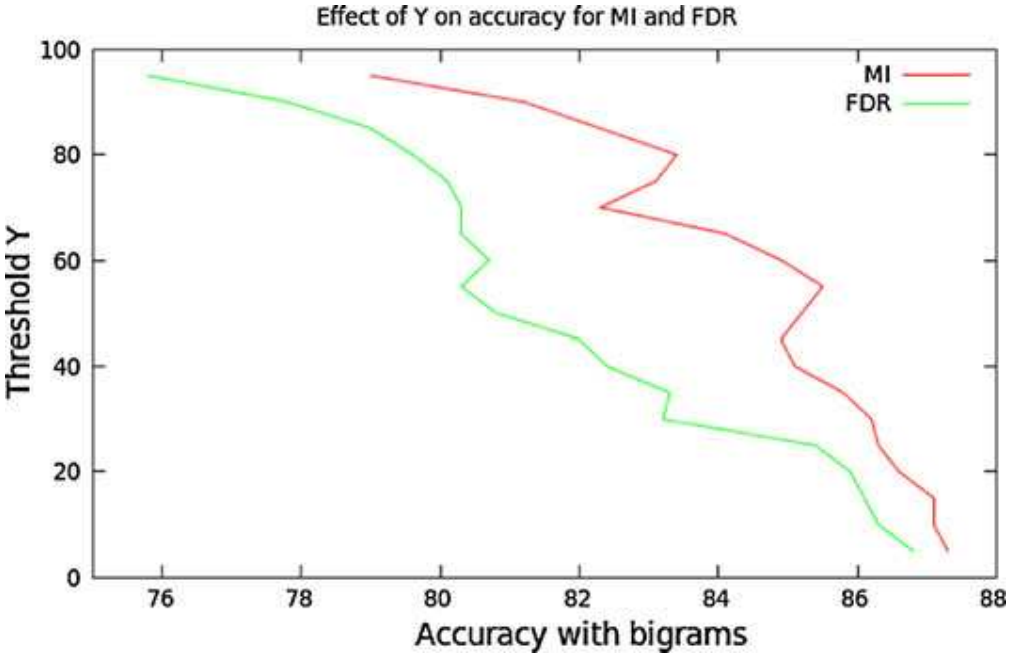
We downloaded the available IMDb archive of the rec, arts, movies, reviews, news, and group. It contains 27,886 unprocessed and unlabeled HTML files that convey opinions of different authors on different movies. Predominantly, it has reviews rated on three different scales: 0–4, 0–5, and grade F to A?. A set of rules are framed to mine the rating patterns from the unprocessed set.

***b. Classifier and evaluation***

We use the SVM classifier implemented in the SVMLight package15 in our experiments, with parameters set to their default values. We focus more on extracting subjective features and representing them as feature vectors to the classifier rather than tuning its parameters.

***B. Estimating parameters***

We start with a very low value of X = 10, i.e., retaining the top 10% of sentences in each review as its subjective excerpt. We then incrementally add 5% in each iteration and examine the increase or decrease in the performance of the classifier.




Tuning the parameter Y for MI and FDR on RSUMM with bigrams as features

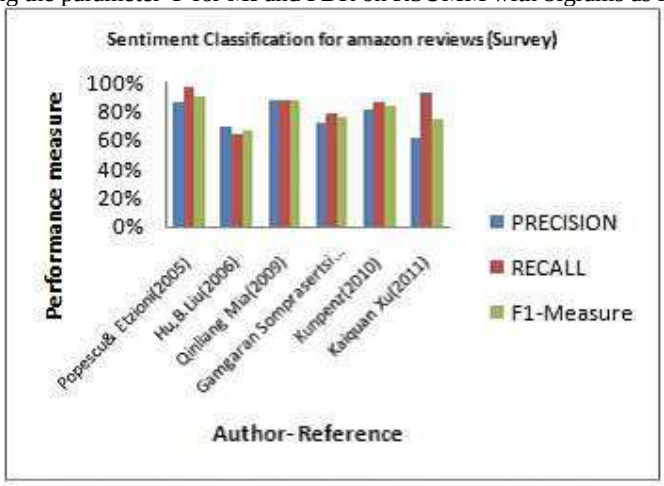



Fig 1. Sentiment Classification for Amazon reviews (Survey)

### *C. Experimental results*

The baseline (BL) in our experiments is using the total review with unigrams (Uni), bigrams (Bi), and their combination (Uni?Bi) as features. We split each review into twoequal halves and carry out experiments using the top half (TH) and bottom half (BH) of the review. This is done to test the general pattern followed by authors in expressing their sentiment. The general pattern in movie review domain is that authors discuss objective information such as plot, characters, and other aspects of a movie at the beginning.

Table 3 reports the accuracy values using DSE and the effect of applying MI and FDR on it. Secondly, we apply MI and FDR on the baseline to verify whether the performance of the sentiment classifier is more sensitive to Y than to X. The results for this experiment are reported.

## VI. CONCLUSION

Sentiment classification is the foundation of sentiment analysis. Sentiment analysis is the process of extracting knowledge from the peoples' opinions, appraisals and emotions toward entities, events and their attributes.

Preliminary tests were done on a subset of the artworks from the art portal ArsMeteo. The proposed approach is particularly suitable to application domains where tags can be interpreted as concise reviews (e.g. artworks, books, movies).

Feature selection techniques have proved vital in the performance of several text categorization tasks, as they enhance the performance of the classification system considerably.

This is on the line of recent approaches which face the challenge of increasing the user involvement in building the Semantic Web an alternative could be to integrate in ArsEmotica the use of automatic techniques.

Finally, future possible uses include the development of emotion-aware search engines and of emotional tag clouds. This would open the way to a plethora of applications, including iOS and Android apps, not only with a cultural flavour (along the lines of the application in the previous section) but also more intrinsically related to leisure.

Our work can be considered as a building block for analyzing sentiment with minimal usage of linguistic resources and no complex patterns. In the future, we need to explore different metrics to extract subjectivity, and conduct experiments. As feature selection methods have proved critical in the performance of classification, we need to explore more novel methods for selecting features.

## REFERENCES

[1] Aigner, W., Miksch, S., M□uller, W., Schumann, H., and Tominski, C. 2007. Visualizing time-oriented data - a systematic view. Comput. Graph.

[2] Aris, A., Shneiderman, B., Plaisant, C., Shmueli, G., and Jank, W. 2005. Representing unevenly-spaced time series data for visualization and interactive exploration.

[3] Ding, X. and Liu, B. 2007. The utility of linguistic rules in opinion mining. In SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and developmentin information retrieval. ACM, New York, NY, USA,

[4] Ding, X., Liu, B., and Yu, P. 2008. A holistic lexicon-based approach to opinion mining. In Proceedings of the international conference on Web search and web data mining.

[5] Hao, M. C., Keim, D. A., Dayal, U., Oelke, D., and Tremblay, C. 2008. Density displays for data stream monitoring. Comput. Graph. Forum

[6] Havre, S., Hetzler, E., Whitney, P., and Nowell, L. 2002. Theme river: Visualizing thematic changes in large document collections. IEEE Transactions on Visualization and ComputerGraphics

[7] Kim, S.-M. and Hovy, E. 2004. Determining the sentiment of opinions. In COLING '04: Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, Morristown, NJ, USA.

[8] Gamgarn Somprasertsri, Pattarachai Lalitrojwong , Mining Feature-Opinion in Online Customer Reviews for Opinion Summarization, Journal of Universal Computer Science, vol. 16, no. 6 (2010), 938-955.

[9] Gang Li , Fei Liu , "A Clustering-based Approach on Sentiment Analysis" ,2010, 978-1-4244-6793-8/10 ©2010 IEEE.

[10] Go,Lei Huang and Richa Bhayani , "Twitter Sentiment Analysis", Project Report, standford,2009.
Go,Lei Huang and Richa Bhayani , "Twitter Sentiment Classification using Distant Supervision", Project Report, Standford,2009.

[11] Guang Qiu , Xiaofei He , Feng Zhang , Yuan Shi , Jiajun Bu , Chun Chen , "DASA: Dissatisfaction-

oriented Advertising based on Sentiment analysis" , Expert Systems with Applications, 37 (2010) 6182–6191.

[12] Hu, and Liu, "Opinion extraction and summarization on the web", AAAI., (2006), pp. 1621-1624.

[13] Hu, and Liu, "Mining and summarizing customer reviews", Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, USA, 2005,pp. 168–177.

[14] Hu, Liu and Junsheng Cheng, "Opinionobserver: analyzing and comparing opinions on theWeb", Proceedings of 14th international Conference onWorldWideWeb, pp. 342-351, Chiba, Japan, 2005