ISSN No. 0976-5697

Volume 6, No. 7, September-October 2015

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

A Survey on Sentiment Classification and Analysis using Data Mining

Ashish Shukla
M.Tech: CSE department. PSIT, Kanpur
Uttar Pradesh Technical University,
Lucknow, India

Shweta Shukla
M.Tech: CSE department. PSIT, Kanpur
Uttar Pradesh Technical University,
Lucknow, India

Abstract: Sentiment analysis is a type of natural language processing that tracks the opinion and mood of the public about the specific product. The volume of information in digital form increases day-by-day as people express their sentiments, also called opinions everywhere mostly on internet as the people now days are much dependent on internet. So the requirement of user opinions analysis is gaining importance day by day. Blogs, micro blogs, review sites, twitter, and other social networks are the most common platforms that are used by people and organizations for posting their views. Researchers has done very immense effort in the field of sentiment analysis and also new opportunities and challenges still arise so even now it is very active and dynamic research area in the field of natural language processing. It is also widely investigated in text mining, data mining and web mining. This paper attempts to evaluate the several techniques used for sentiment analysis. The goal of this survey is to give an in-depth introduction to this fascinating problem and to present a comprehensive study of all important research using data mining and the latest developments in the field.

Keywords: Opinion Mining, NLP, Sentiment Analysis, Information Retrieval, Naïve Bayes Algorithm

I. INTRODUCTION

Sentiment analysis or opinion mining is used to build a system that collect and analyze feedbacks of customers about the specific product or service. It tracks the public feelings and mood about a certain product or service they are using. People give their feedbacks and share their opinions in blogs, review sites and other social networking sites like Twitter and Face book. Sentiment analysis is very important and crucial for market competitors. It helps them in their decision making process. They can identify which particular product or which product feature is more suitable for particular geographic or demographic region. Sentiment classification has many applications in several fields and across different domains like business and e-commerce industry, government intelligence, review-related websites and as a sub-component technology. It can be used to classify the product reviews into positive and negative class. This is very helpful for the new customers in gaining the overall idea of what other existing customers are saying about that product so that they can decide whether the product should be bought or not.

One of the major applications of sentiment analysis is Text Categorization. It is the process of classifying written text documents into some categories or classes from a pre-defined training dataset. It is widely used in many applications related to Natural Language Processing and has gained considerable attention in recent years from researchers as well as the academic and industry developers. There are many opportunities and new challenges are arising continuously in the field of sentiment analysis. There are some basic problems are encountered when we talk about sentiment classification. For example a particular word may have ambiguous appearance that means sometimes it behaves like positive word and sometimes behaves like negative word depending upon the situation. Also traditional text processing process says that small difference among the text documents do not change the overall meaning very much but in sentiment analysis process but it has to be kept in mind that customers express their sentiments in different ways and not always in a same way. Moreover most of the comments or reviews made by people have both positive and negative statements and there may be a contradiction in customer comments.

The remaining paper is described in the following sections; section II describes some of the previous studies done on sentiment analysis, section III describes several data sources used for sentiment analysis, section IV describes text levels on which sentiment analysis is applied, section V illustrates some well-known sentiment classification techniques, and finally section VI summarizes conclusion and the future work.Type Style and Fonts

II. PREVIOUS STUDIES

Many Scholars and Researchers have studied various aspects of sentiment analysis and opinion mining, some of which are described here. Jalaj S. Modha, Prof & Head Gayatri S. Pandi Sandip J. Modha, **Automatic Sentiment Analysis for Unstructured Data**, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 12, December 2013[20] discussed about exiting methods, approaches to do sentimental analysis for unstructured data which reside on web. They proposed new approach to classify and handle subjective as well as objective statements for sentimental analysis. According to Bing Liu, "**Sentiment Analysis and Opinion Mining",** Morgan & Claypool Publishers, May 2012. [19], Opinions and its related concepts such as sentiments, attitudes, and emotions are the subjects of study of *sentiment analysis*. Since the Web became widespread, the amount of electronically available information online, especially news archives, has proliferated. **K. Bun and M. Ishizuka, "'Topic extraction from news archive using TF*PDF algorithm "** propose an information system that will extract main topics in a news archive on a weekly basis. By obtaining a weekly report, a user can know what the main news events were in the past week. In general, related research on subject

identification is classified into two types. First one is term weighting method to extract useful terms that is relevant to collected documents and modeled also. Second is TF-IDF mostly used for term weighting in Natural language processing and information extraction process.[16] Fred Popowich**, Using Text Mining and Natural Language Processing for Health Care Claims Processing**, SIGKDD Explorations. Volume 7, Issue 1 - Page 59 [17] described that the application makes use of a natural language processing (NLP) engine, together with application-specific knowledge, written in a concept specification language. Using NLP techniques, the entities and relationships that act as indicators of recoverable claims are mined from management notes, call centre logs and patient records to identify medical claims that require further investigation. Text mining techniques can then be applied to find dependencies between different entities, and to combine indicators to provide scores to individual claims. Claims are scored to determine whether they involve potential fraud or abuse, or to determine whether claims should be paid by or in conjunction with other insurers or organizations. Dependencies between claims and other records can then be combined to create cases. Arti Buche, Dr. M. B. Chandak, Akshay Zadgaonkar, **OPINION MINING AND ANALYSIS: A SURVEY,** International Journal on Natural Language Computing (IJNLC) Vol. 2, No.3, June 2013 [18] is focusing on the area of Opinion Mining also called as sentiment analysis. They surveyed and analyzed various techniques that have been developed for the key tasks of opinion mining. They have provided an overall picture of what is involved in developing a software system for opinion mining on the basis of our survey and analysis. Classifying entire documents according to the opinions towards certain objects is called as sentiment classification. One form of opinion mining in product reviews is also to produce feature-based summary. To produce a summary on the features, product features are first identified, and positive and negative opinions on them are aggregated. Features are product attributes, components and other aspects of the product. The effective opinion summary, grouping feature expressions which are domain synonyms is critical.

III. DATA SOURCES

The major criterion for the improvement of the quality services rendered and enhancement of deliverables are the user opinions. Blogs, micro blogs and review sites serve as rich data sources for sentiment classification and analysis.

Blogs

A Blog is a webpage that contains information about someone's activities or interests. People can read a blog and can write their own opinion about what it contains. Usually blogs are updated frequently. People exchange their views with one another on the topics they want to discuss on a blog. There are millions of messages are posted at a time and these blogs are used for sentiment analysis. [13]

Review Sites

There are plenty of websites are available on the internet in which thousands of consumers are generating reviews for products and services they are using. These reviews play important role in decision making for the new user about what to purchase and what to not. In sentiment analysis and classification customer reviews data is needed that is available on the different websites like www.reviewcentre.com (product reviews), www.fonearena.com (mobile reviews), www.flipkart.com (product reviews), in which thousands of product reviews are available commented by consumers. [14]

Micro-blogs

Micro blog is a kind of blog that enables users to broadcast short text messages or media i.e. pictures, video, or sounds to other users of the service. Social networking sites, like Twitter or Face book are the most commonly and widely known examples of micro blogs. Sometimes these Twitter messages express sentiment that can be assumed as the data source for sentiment classification and analysis. [15]

IV. DIFFERENT LEVELS OF SENTIMENT ANALYSIS

The information is collected from online reviews for sentiment analysis. The sentiment analysis can be performed at one of the following three levels:

Document Level Sentiment Classification: In this type of classification informative text has to be extracted and analyzed for inferring sentiment of the whole document. The document is taken as a whole and labeled to a particular class. For this type of classification, Supervised Learning approaches are used. [11]

Sentence Level Sentiment Classification: It is also known as clause level classification. In this type of classification, each sentence is assumed as an entity, then analysis is done on individual sentence and finally their result is summarized to obtain the overall sentiment of the document. This sentiment classification is a fine-grained level than document level sentiment classification. [11]

Feature Level Sentiment Classification: The feature level sentiment classification is much more pinpointed process to sentiment analysis. Product features are defined as product attributes or components. Analysis of these features for sentiment classification of the document is known as feature based sentiment analysis. In this approach positive or negative opinion is identified from the already extracted features. [12]

V. SENTIMENT CLASSIFICATION TECHNIQUES

Generally sentiment analysis can be performed at the following 3 levels: the document level, sentence level, aspect or attribute level [11] [12]. Many techniques and methods of natural language processing are being used here in sentiment analysis more specifically for sentiment classification at the document level. So sentiment detection therefore shares information, knowledge and many properties with information retrieval and natural language processing systems for example text mining, text search predicative analysis, effectiveness measures etc. This section provides brief details of the supervised, unsupervised and semi-supervised learning and other algorithms used in the experiments.

1. SUPERVISED LEARNING

Since early 2000, researchers have been studying about Machine learning, also known as supervised learning and using this they derived opinions from feedbacks and reviews posted online [1]. In this approach, a set of labeled training documents of each class is used by a learning algorithm to build a classifier. Several machine learning techniques have

been applied to sentiment classification. The most widely used supervised learning techniques for sentiment classification for product reviews are Naïve Bayes(NB) Classification, Maximum Entropy(MaxEnt), Support Vector Machines(SVM), Neural network, Multi-Layer Perceptron (MLP), Decision tree. These techniques need training data to perform and for this dataset of labeled opinion words are needed. This section provides brief details of the machine learning algorithms used in the experiments.

A. Naive Bayes (NB): It is probabilistic classifier that uses Bayes Theorem, Which finds the probability of an event given the probability of another event that has already occurred. According to bayes theorem, the probability that we want to compute P(X | Y) can be expressed in terms of probabilities P(X), P(Y | X) and P(Y) as,

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

Naïve bayes classifier is well suited when the dimensionality of inputs is high. It gives more accurate and efficient results for linearly separable cases and even performs well for non-linearly separable cases [3]. Naïve Bayes algorithm can be represented using figure 1:

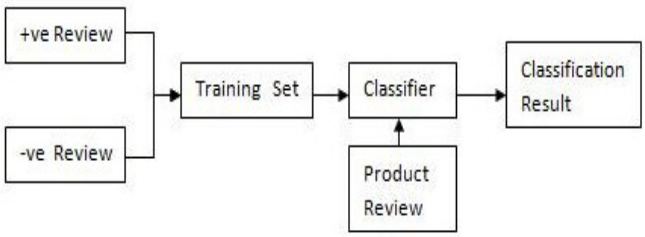

Fig1: Naïve bayes classification

Major advantage of Naïve Bayes Classification is easy to interpret and it has efficient computation.

B. Multi-Layer Perceptron (MLP): An MLP also known as Artificial Neural Network (ANN) can be considered as network of neurons called perceptrons. The perceptron computes a single output from multiple inputs.MLP is also known as feed forward networks and can have one or more hidden layers between input and output layer. The MLP networks can be used for both supervised and unsupervised learning process. [2]

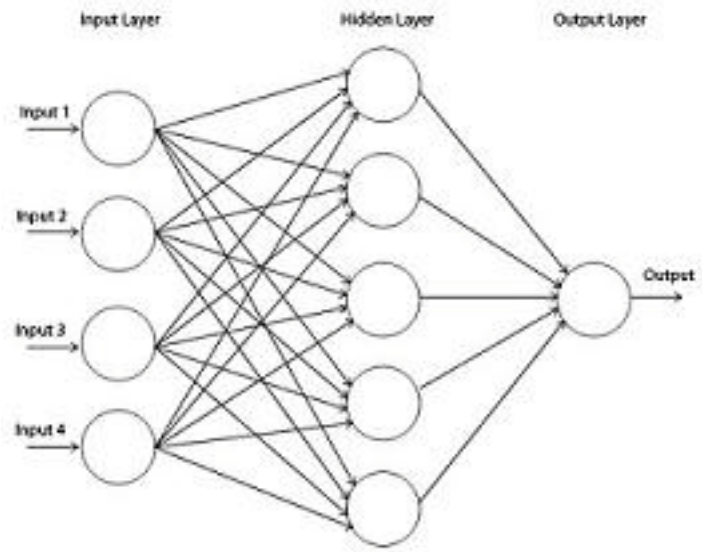

Figure 2:MLP

The above architecture has the following properties: 1.There is no connections within a layer, 2.There is no direct connections from input to output layers, 3.The layers are fully connected, 4.Generally there are more than 3 layers, 5.It not necessary that the no. of input units are equal to the no. of output units, 6.No.of hidden units in each layer can be more or less than input or output units.

The MLP network should have minimum three hidden layers for any valid representation and such a network takes much time for its training process.MLP is the most used type of neural network algorithm and having huge number of applications. It is capable of modeling complex functions. It is very good at ignoring irrelevant inputs and noise and it can be used even if a few knowledge available about the relationship of the function to be modeled.

C. Support Vector Machine (SVM): This classifier constructs N-dimensional hyper plane which separates data into two categories. SVM models are closely related to a Neural Network. SVM takes the input data and for each input data row it predicts the class to which this input row belongs. SVM works for two class problems and is a non probabilistic binary linear classifier [4]. The concept of SVM algorithm is based on decision plane that defines decision boundaries. A Decision plane separates group of instances having different class memberships. For example, consider an instance which belongs to either class Circle or Diamond. There is a separating line (figure 3) which defines a boundary. At the right side of boundary all instances are Circle and at the left side all instances are Diamond.

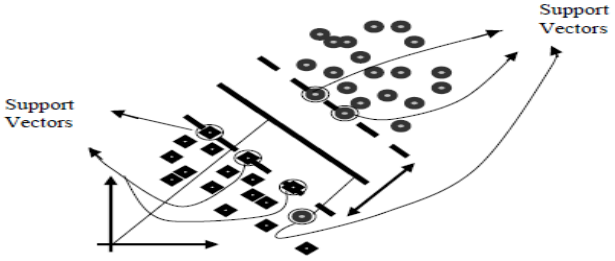

Figure 3. Principle of SVM

Support Vector Machine Method performs very well on experimental results and has low dependency on data set dimensionality.

2. SEMI-SUPERVISED LEARNING

Reading and labeling documents (often manually) requires more cost and effort and moreover classification is a time-consuming and boring task that has to be done by humans. So to overcome the problem of manual labeling, *semi-supervised learning* is studied. Semi-supervised learning generates a suitable function or classifier in which both labeled and unlabelled examples are combined [23].

It is easier to find unlabeled (also called unclassified) than labeled (also called classified) documents in many classification tasks. In such cases, we can improve the results obtained by the classifier by using both, the labeled documents as well as the unlabeled documents that are available. This task is called semi-supervised classification. Semi-supervised Machine Learning methods can train with a small set of labeled examples and use unlabeled examples to improve their model of the data. [21][22]

3. UNSUPERVISED LEARNING

Unsupervised Learning tries to find the hidden structure in unlabeled data. That is why it does not require any prior training in order to analyze the data. Instead of that, it tries to measure how far a particular word is tending towards positive and negative sentiment. Clustering algorithm, expectation-maximization algorithm, matrix factorization, principal component analysis and many others are the common examples of unsupervised learning. Much of the research in unsupervised sentiment classification makes use of lexical resources available. Kamps et al [5] used lexical relationships in sentiment analysis and classification. Andrea Esuli and Fabrizio Sebastian [6] proposed semi-supervised term classification for determining the orientation of subjective terms. When the review have not enough contextual information to determine the actual sentiment, Chunxu Wu[7] proposed a method in which contextual information present in other reviews about the same topic is gathered and analyzed, then by using semantic similarity among them, one can judge the orientation of that sentiment. An unsupervised learning algorithm by extracting the sentiment phrases of each review by rules of part-of-speech (POS) patterns was investigated by Ting-Chun Peng and Chia-Chun Shih [8]. For each unknown sentiment phrase, they used it as a query term to get top-N relevant snippets from a search engine respectively. Next, by using a gathered sentiment lexicon, predictive sentiments of unknown sentiment phrases are computed based on the sentiments of nearby known sentiment words inside the snippets. Gang Li & Fei Liu [9] developed an approach based on the k-means clustering algorithm. This approach used the phenomenon of TF-IDF (term frequency – inverse document frequency) weighting applied on the raw data. After that an efficient clustering algorithm is applied to derive best clustering results. Chaovalit and Zhou [10] compared two approaches namely; Semantic Orientation approach and N-gram model machine learning approach. They applied both of these on movie reviews.

4. FEATURE BASED SENTIMENT CLASSIFICATION

As the amount of reviews and feedbacks given by people on the internet increases, Sentiment analysis is gaining importance and has become a versatile topic in natural language processing. Here is some feature definition and feature selection strategies for sentiment classification are discussed.

Feature Definition

Words and Stems: Though a document may certainly be represented by the raw words existing in it, there is a classical method in information retrieval system that stem the words to their morphological roots. Stemmed feature vectors are normally smaller in size. These vectors aggregate across occurrences of variants of a particular given word. Stemming has found successful in both information retrieval and text mining, for example, it has been shown that stemming produces mixed results on different types of datasets.[30] They conclude that “corpus of reviews is highly sensitive to minor details of language, and these may be glossed over by the stemmer”.

Binary and Term Frequency Weights**:** In information retrieval term frequency (TF) weights approach is used to show the relative importance of features in document representations. On the othr hand, some other researcher have shown that binary weighting (1 if the wordoccurs in the document, 0 otherwise) is more beneficial for sentiment classification .The standard IR weighting schemes in Sentiment Analysis is studied and , [24] found that using binary features is better than raw term frequency, though a scaled TF version28 performs as well as binary.

Negations: Negations are often included in stopword lists, and hence are removed from the text analysis. But when these are combined with other words, negations reverse the polarity of words. Because negations may affect the sentiment classification, SA researchers have tried incorporating them into the feature vector. [25] use a heuristic to identify negated words and create a new feature by appending NOT- to the words (for example, a phrase “don’t like” results in feature NOT-like).

N-grams: These are ordered sets of words. Negation phrases discussed earlier are the special case of n-grams.The benefit of using n-grams instead of single words as features is that we are able to capture some dependencies between the words and the importance of individual phrases. In a study of subjective text fragments, [26] show that it is beneficial to use higher order n-grams(upto 6) in sentiment classification task.

Feature Selection

Frequency-Based Selection: In text modeling, it is often the practice to remove words which appear rarely in the corpus. These are presumed to be perhaps misspellings, that do not help in generalization during classification. On the other hand, words that occur only once in a given corpus have been found to be high-precision indicators of subjectivity [27]. Thus rare terms may serve an crucial role in sentiment classification.

Mutual Information Based Selection: The performance of the classifier may also be improved by removing some of the less useful features. One of the common feature selection measurements is expected Mutual Information. Usually the features are scored by the expected MI and top several are taken as the most useful in classification.

Part of Speech-Based Selection: Particularly for Sentiment Analysis, certain POS have been determined to be more useful in classification tasks. For example,[28] showed that using adjectives and adverbs works better than using adjectives alone, also use verbs for sentiment classification. If indeed adjectives are important factors in predicting sentiment

polarity, limiting the feature space to only these may improve classifier performance by removing less useful words.

Lexicon-Based Selection: Sentiment-annotated lexicons may be used for feature selection. Popular lexicons are the extensions of WordNet (http://wordnet.princeton.edu/), a large lexical database of English. for example, [29] described SENTIWORDNET, a lexical resource in which each WORDNET synset s is associated to three numerical scores Obj(s), Pos(s) and Neg(s), describing how objective, positive, and negative the terms contained in the synset are.

VI. CONCLUSION AND FUTURE WORK

In the above survey different methods and techniques for opinion mining are presented. Every method has its own importance, some pros and cons. These methods have been applied in various types of dataset. A comparative study has been evaluated on the basis of above strategies and has been find that SVM performs better than others. According to the situation, a particular method is used and can be applied to different areas. From our point of view Naïve Bayes is performs well for text based classification and SVM for biological interpretation. In future we will be finding out the best result of sentiment analysis by applying and combining these methods on different datasets like social networking reviews.

VII. REFERENCES

[1] "Sentiment classification using machine learning techniques." Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79–86.

[2]Towards Enhanced Opinion Classification using NLP Techniques, IJCNLP 2011, pages 101–107, Chiang Mai, Thailand, November 13, 2011

[3] Qiang Ye, Ziqiong Zhang, Rob Law, "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches", Expert Systems with Applications 36 (2009) 6527–6535.

[4] Rui Xia, Chengqing Zong, Shoushan Li, "Ensemble of feature sets and classification algorithms for sentiment classification", Information Sciences 181 (2011) 1138–1152.

[5] Kamps, Maarten Marx, Robert J. Mokken and Maarten De Rijke, "Using wordnet to measure semantic orientation of adjectives", Proceedings of 4th International Conference on Language Resources and Evaluation, pp. 1115-1118, Lisbon, Portugal, 2004.

[6] Andrea Esuli and Fabrizio Sebastiani, "Determining the semantic orientation of terms through gloss classification", Proceedings of 14th ACM International Conference on Information and Knowledge Management,pp. 617-624, Bremen, Germany, 2005.

[7] Chunxu Wu, Lingfeng Shen, "A New Method of Using Contextual Information to Infer the Semantic Orientations of Context Dependent Opinions", 2009 International Conference on Artificial Intelligence and Computational Intelligence

[8] Ting-Chun Peng and Chia-Chun Shih , "An Unsupervised Snippet-based Sentiment Classification Method for Chinese Unknown Phrases without using Reference Word Pairs", 2010 IEEE/WIC/ACM International Conference on Web Intelligence and intelligent Agent Technology JOURNAL

[9] 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.

[10] Kennedy and D. Inkpen, "Sentiment classification of movie reviews using contextual valence shifters," Computational Intelligence, vol. 22, pp. 110–125, 2006.

[11] V. S. Jagtap and Karishma Pawar, Analysis of different approaches to Sentence-Level Sentiment Classification, International Journal of Scientific Engineering and Technology (ISSN : 2277-1581) Volume 2 Issue 3, PP : 164-170 1 April 2013

[12] Zhongwu Zhai, Bing Liu, Hua Xu and Hua Xu, Clustering Product Features for Opinion Mining, WSDM'11, February 9–12, 2011, Hong Kong, China. Copyright 2011 ACM 978-1-4503-0493- 1/11/02...$10.00

[13] Singh and Vivek Kumar, A clustering and opinion mining approach to socio-political analysis of the blogosphere, Computational Intelligence and Computing Research (ICCIC), 2010 IEEE International Conference.

[14] G. Vinodhini and R.M. Chandrasekaran, Sentiment Analysis and Opinion Mining: A Survey, Volume 2, Issue 6, June 2012 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering

[15] Alexander Pak and Patrick Paroubek, Twitter as a Corpus for Sentiment Analysis and Opinion Mining

[16] K. Bun and M. Ishizuka "Topic extraction from news archive using TF*PDF algorithm" In Proceedings of Third International Conference on Web Information System Engineering.

[17]. Fred Popowich, Using Text Mining and Natural Language Processing for Health Care Claims Processing, SIGKDD Explorations. Volume 7, Issue 1 - Page 59

[18] Arti Buche, Dr. M. B. Chandak, Akshay Zadgaonkar,OPINION MINING AND ANALYSIS: A SURVEY, International Journal on Natural Language Computing (IJNLC) Vol. 2, No.3, June 2013

[19] Bing Liu. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012

[20] Jalaj S. Modha, Prof and Head Gayatri S. Pandi Sandip J. Modha, Automatic Sentiment Analysis for Unstructured Data, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 12, December 2013

[2] Blum, A., and Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. *COLT-98.*

[22]Nigam, K., McCallum, A., Thrun, S., and Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39

[23] Bing Liu. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012.

[24] Paltoglou, G., and Thelwall, M. 2010. A study of information retrieval weighting schemes for sentiment analysis. *Proc. of ACL* 1386–1395

[25] Das, S., and Chen, M. 2001. Yahoo! for Amazon: Extracting market sentiment from stock message boards. *Proc. Of APFA.*

[26] Hang Cui, Vibhu Mittal, and Mayur Datar. Comparative experiments on sentiment classification for online product reviews. National Conference on Artificial Intelligence (AAAI), 21(2), 2006.

[27] Wiebe, J. M.; Wilson, T.; Bruce, R.; Bell, M.; and Martin, M. 2004. Learning subjective language. *Computational Linguistics* 30.

[28] Benamara, F.; Cesarano, C.; Picariello, A.; Reforgiato, D.; and Subrahmanian, V. 2007. Sentiment analysis:

Adjectives and adverbs are better than adjectives alone. *Proc. Of ICWSM*.

[29] Esuli, A., and Sebastiani, F. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. *Proc. Of LREC*.

[30] Dave, K.; Lawrence, S.; and Pennock, D. M. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. *Proc. of WWW*