International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

# Sentiment Analysis and Challenges Involved: A Survey

**Archana Sonagi[1], Deipali Gore[2]**

Savitribai Phule Pune University, Progressive Education Society's, Modern College of Engineering,
Shivaji Nagar, Pune-411005, Maharashtra, India

**Abstract**: ***Web is a global fence with enormous opportunities for people to share their thoughts and to know their opinions about almost anything in the world. Influence of what other people think has always been a prime part of ones life from generations for a better decision making. The abundant amount of opinions or views expressed by internet users today not only consists of a wealth of intelligent information but also gains monetary benefits, and the urge to be benefited from this knowledge source has led to emergence of a field known as Sentiment Analysis. Opinion Mining is a computational study of opinions which applies Natural Language Processing techniques and computational linguistics which extracts the view of users from various sources on web and analyzes them to identify their polarity. The study investigated briefs about the framework of a Sentiment Analysis system used in general, the various supervised and lexicon based approaches used for the classification, the challenges that are confronted in building an efficient system and lastly, it covers the various applications of Sentiment Analysis.***

**Keywords:** Natural Language Processing (NLP), Sentiment Analysis, Opinion Mining**,** Classification Techniques, Challenges involved.

## 1. Introduction

Internet today has changed the way people interact with each other and know each other. Web in general has presented us a platform in the form of Social Media, Forums, and Blogs etc. to raise our opinions publicly and to know their feedbacks about a certain topic from people almost anywhere in the world. It has changed the outlook of people towards the internet from just being a "Read Only" platform to "Read-Write". The users hunger for and dependence on online advice and recommendations has drawn interest of researchers to research in this area. This need to analyze the thoughts of people and to gain wealth of information from it has led to the emergence of the field of Opinion Mining.

Sentiment Analysis is also popularly known as Opinion Mining. It can be defined as a sub discipline of Natural Language Processing and Computational Linguistics mainly concerned with the emotion, thought, or a mood expressed by a reader in any document. Here, the former term signifies evaluation of information when extracted and the latter denotes drawing or outing of subjective information from a text corpus or reviews [1, 2].

Web is a huge repository of both, structured and unstructured form of data. Building a system to explore user's opinions available on the web in the form of reviews, Blogs, Forums, and Social Media etc. is a very crucial task. The richness of the language and the freedom to express in a free form language makes it even more challenging. Also the fact, that the presence of biased or manipulated reviews posted with a certain intention by the author, raises a question of authenticity of reviews in people's mind about to what extent should one trust them. This way of manipulating reviews is known as Opinion Spamming and the ones those perform these acts are known as Spammers. Other than sentiment classification the other major areas of research interest include Opinion Summarization which summarizes only the features of the product that are mined and Feature based Sentiment Classification which considers the specific features of certain objects.

The Question that arises is why is there a need for Sentiment Analysis? Does web really contain such sentiment related information? If Yes, then where and how much. There is a great volume of opinionated text available on the web which grows on escalating everyday in the form of Review Sites, Wikis, Micro Blogging, Social Media (Facebook, Twitter, YouTube) and many more. To derive benefits from this data arises a need to collect, organize and analyze the information to be available of use to all. The Analysis includes huge monetary benefits at both producer and Consumer level.

## 2. The Levels of Sentiment Analysis

**1) Document level sentiment Analysis**
Determines the overall sentiment of a given review without considering the individual aspects provided in the document. In such a classification it is assumed that the opinion expressed is on a single entity and has a single opinion holder. Such documents are considered to be opinionated. Many machine learning techniques like NB, SVM have been used for detecting the polarity of a document.

**2) Sentence level sentiment Analysis**
Sentence level classification deals with classifying polarity at sentence level in a document. Information in a sentence can be of two types, Subjective i.e. containing relevant information or Objective i.e. containing a neutral opinion or just a fact. Non opinionated text is then eliminated and classification is performed on the subjective part of the sentence.

**3) Feature level sentiment Analysis**
Feature level Sentiment classification is done with a very fine grained analysis of every attribute. Sentiments in a review may be expressed with respect to different features. A review with a overall positive opinion does not necessarily mean that

the author likes everything about the object or vice-versa. It firstly identifies and extracts the features, determines the polarity and then groups the feature synonyms.

## 3. Framework of Sentiment Analysis Approach

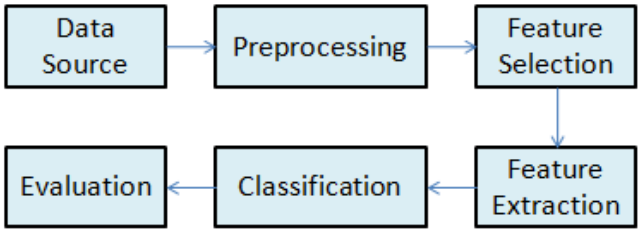


**Figure 1:** General Framework

### 3.1 Data Sources

The sources mentioned below are widely used by people for finding opinions and giving recommendations for a particular application.

**1) Blogs**
Blog pages contain expression of one's opinion related to any topic say an event, issues or a service written in a informal or conversational style. Blogs are a regularly updated, effective and fast way to share your news and views with others. The popular blogging providers are WordPress.com, LiveJournal, and Blogger.

**2) Review Sites**
They are the direct sources to know the users opinion about businesses, products or services for decision making in a more clear and summarized way highly influencing consumers choices. Some of the well known sites are Angie's List, Yelp, Glassdoor, TripAdvisor etc.

**3) Social Media**
These are the highly interactive platforms through which individuals can share, co-create, discuss and modify user generated content. Some of the strongest media used by people over the world today are Facebook, Tumblr, YouTube containing tremendous amount of knowledge literally about any domain.

### 3.2 Preprocessing

Real world data is often incomplete and inconsistent and is likely to contain many errors. In this phase the raw data is processed.

**1) Tokenization**
It is the process of splitting a stream of text into words, phrases or some meaningful elements called tokens with the help of boundaries between words marked by special delimiting characters such as spaces, punctuations and symbols. This process is also known as word extraction, word segmentation or lexical analysis.

**2) Stop words Removal**
These are the very commonly occurring words in text having a very low discriminative value, serving only the syntactic meaning. It involves creating a list of stop words and then scanning the document so that word appearing in the stop list is removed.

**3) Stemming**
It is a process to reduce a word to its stem or root word. It is widely used in Information Retrieval to increase the recall rate.

**4) Case Normalization**
The text published contains both the uppercase and lowercase characters; this process converts the entire text in either Uppercase or Lowercase.

### 3.2 Feature Selection

Feature Selection is a method which reduces both the data and the computational complexity. The set of features that describe a particular object has a greater impact on the classification of objects and thus should be chosen smartly. Data along with useful features usually, may also contain redundant features i.e. the ones providing no extra information or irrelevant features providing no useful information. Thus it becomes a critical task to select a more appropriate subset of features. The most efficient subset would be the one which minimizes error rate making the task more effective and accurate.

**Table 1:** Feature Selection Techniques

| *Sr. No.* | *Techniques Used* |
|---|---|
| 1 | Information Gain |
| 2 | Odd Ratio |
| 3 | Document Frequency |
| 4 | Mutual Information |
| 5 | Point-Wise Mutual Information |

### 3.3 Feature Extraction

Feature Extraction creates the new features from the functions of original features. When the input data is too large and contains so much of redundant data, this reduction technique is used to extract a reduced set of the most relevant features while still describing the data with sufficient accuracy. Feature Extraction can be combined with Dimensionality reduction using techniques like PCA, Term Presence, Term Frequency, Standard Deviation and TF-IDF and by elimination of low level or unwanted linguistic features. The three main methods of feature extraction are Filter Techniques, Wrapper techniques and Embedded techniques. Filter methods select the best of features based on intrinsic criterion such as distance measures, Wrapper methods selection is based on generation and evaluation of different subsets in space of states and embedded method looks for an optimal subset of features via a search in hypotheses and space of feature subset [3].

### 3.4 Classification

**1) Machine Learning Based Approaches**

In any machine learning based classification approach two sets of documents are required: Training and a test set. A training set is used for learning different characteristics of a document by a classifier and a test set validates the performance of an automatic classifier. There are a number of machine learning techniques adopted for the classification some of which are SVM, NB, NN etc.

Support Vector Machine is a Machine based learning method in which a decision boundary which separates the two classes with maximum margin between the training points of two classes is found. M. Rushdi Saleh et al. [4] Applied SVM for multi domains such as blogs and product reviews using different weighting schemes which have provided a accuracy of 91.51%.Rodrigo Moraes et al. [5] in his comparative study between SVM, ANN, NB, ME have experimented with movie reviews with this classification techniques, in which he used unigrams, bigrams and information gain for feature selection and TF-IDF as a weighting mechanism. His experiments conclude that ANN performs better than SVM in case of both balanced and unbalanced data. IG helps in removing the noisy data and reduces the running time of SVM.

Alexander Pak and Patrick Paroubek [6] presented a method for automatic collection of corpus from twitter in which a tree tagger for POS tagging is used and then the difference in distributions of positive, negative and neutral sets is observed. The classifier is based on multinomial naïve Bayes classifier that uses N-gram and POS tags as features. The use of salience as filter gives better result than the entropy for n-grams. Rui Xia et al. [7] use Naïve Bayes to estimate the probabilities of categories given a test document by using the joint probabilities of words and categories. The simplicity of NB makes the process efficient.

A. Nisha Jebaseeli and E. Kirubakaran [8] proposed a Neural Network Classification with m-learning reviews in which they used Singular Value decomposition to score the opinion words, Information Gain for feature selection and an improved input output Feed Forward Neural Network for classification. The results proved to be more efficient than other neural network algorithms.

**2) Semi supervised Approach**

Guang Qiu et al. [9] analyzed problems related to opinion lexicon i.e. list of opinion words indicating notions of sentiments and opinion target extraction i.e. the entities and their attributes on which opinion have been expressed. The semi supervised method employed uses opinion words as seeds in initial opinion lexicon and bootstrapping is then initiated on it. Double propagation method is used as information is propagated back and forth between opinion words and targets. Suke Li [10] proposed a semi supervised Sentiment classification method for web consumer reviews. The method is based on the co-training framework with three basic sentiment classifiers of which the first is constructed with common unigram features; the second is based on the extracted opinion words from subjective views. While, the remaining are trained for the third classifier. The results are effective and said to outperform the self learning SVM method.

**3) Lexicon Based Approaches**

Unsupervised Sentiment Classification does not require prior training in order to mine the data. It measures how far a word is inclined towards positive or negative notion. A list of words or phrases is called a lexicon. X. Glorot et al. [11] in his work has focused on the domain adaptation problem. As the reviews can span so many different domains it is difficult to gather annotated training data for all of them. His paper has demonstrated a deep learning approach based on Stacked Denoising Auto-Encoders with sparse rectifier units that perform an unsupervised feature extraction highly beneficial for domain adaptation of sentiment classifiers followed by the training of linear classifiers. The results are more improvised than a purely supervised alternative.

Gang Li and Fei Liu [11] introduced a clustering based approach where TF-IDF is used as a weighting mechanism, voting mechanism to obtain more stable clustering results and the term scores are imported for improving the performance. It is said to produce accurate clusters within shorter time along with more balanced performance in terms of accuracy, efficiency and least human participation. Gang Li and Fei Liu [12] have further extended his research to produce accurate results without any linguistic knowledge, human participation or training time. The two aspects that are focused on are opposite and non-opinion content processing techniques with the use of distance measurement methods and modified voting to conduct fine grained analysis. The achievements are more efficient then supervised learning approaches.

### 3.5 Evaluation

The most commonly used measures of evaluation are Precision, Recall, F-measure and Accuracy. Precision gives the fraction of retrieved instances that are relevant while recall is the fraction of relevant instances that are retrieved. F-measure is a harmonic mean of both precision and recall. Accuracy gives a measure of how close a value is to the true (actual) value.

## 4. Challenges Involved

**1) Negation**

Correctly determining the valence of a text is equivalent to the success or failure of the automatic processing. The features for negation modeling are organized in three groups:

- Negation features
- Shifter features
- Polarity modification features

Negation features relate directly to the position of negating expression as to whether it appears in a fixed window of four words preceding a subject or in a predicate. Shifter features are the binary features checking the presence of different types of polarity shifters. General polarity shifters invert the polarity of an expression. Polarity modification features either modify other expressions or themselves get modified [13].

**2) Domain Dependent**
Sentiments are expressed differently in different domains. The classifiers trained to classify a certain word as being positive in one domain may not prove to be effective if the same word is used to express a view or thought in different domain. Researches to resolve this issue of Domain Dependence still have a large scope of improvement.

**3) Mixed Nature of Opinions**
Web has provided us with platforms where one can be informal in the way they express their view such as blogging. The same sentence may contain a positive as well as a negative opinion about a certain thing which at times may not be easily interpreted by humans too making it even complicated for machines to interpret them.

**4) Entity Identification**
An opinion may consist of more than one entity. They are definite noun phrases that may refer to individuals, businesses, organizations, Dates etc. It is very necessary thus to identify the entity towards which the author is mainly directed.

**5) Co-Reference Resolution**
In linguistics, when two or more expressions in a text refers to the same person or a thing then a co-reference is said to occur. In such sentences one part is usually an antecedent and the other part is an anaphor (an abbreviated form). This task basically helps in providing more information in the information retrieval tasks.

For instance consider, "The music was so loud that it couldn't be enjoyed." The pronoun "it" in English has many uses, but generally refers to an inanimate object.

## 5. Applications

Sentiment Analysis is widely used today in many diverse fields. Organizations ranging from small to big use sentiment analysis in business intelligence to decide on market strategies, to discover the consumer trends, to know about how satisfied their customers are with their services etc. Politicians find it useful to post their views or strategies; they get to know about the demand of the public or their views as of what are their expectations, who are their main followers and thus can modify their strategies accordingly. To know of biased or manipulated comments in news sources. Ad placements use them to find an appropriate place to get placed so that they can attract the interested audience or in case may remove the ads if they find it ill suited for the target audience. It also has many uses in Psychology to augment investigations and Sociology for idea propagation, Law/Policy making etc. Thus with the growing advancements in technology and the will to prosper sentiment analysis is being used smartly in every possible field.

## 6. Conclusion

There has been an extensive research in this field in the recent years leading to represent Sentiment Analysis as a separate research field. The applications of this are varied and benefit not only the small scaled or big organizations but also in areas related to everyday life such as education, tourism, politics etc. The paper briefs a survey on the general techniques followed by any SA system, the challenges that are yet to be resolved to make an efficient and fully automated system, the machine learning and the lexicon based approaches implemented by the researchers, and the applications. The field of opinion mining is highly context dependent and thus there is no general technique which fits all, each has some merits and demerits which when used in combinations may prove to be more accurate. The richness of the language and the freedom to express views in a more informal way complicates the interpretation by the machines while the other major issue that needs to be focused is evaluating the trustworthiness of the opinion and its source.

## 7. References

[1] Alireza Yousefpour, Roliana Ibrahim, Haza Nuzly Abdull Hamed, "A Novel Feature Reduction Method in Sentiment Analysis", IJIC, pp. 34-40, 2014.
[2] B. Pang and L. Lee, and S. Vaithyanathan, "Thumbs Up?: Sentiment Classification using Machine Learning Techniques", in Proceeding of the ACL-02 Conference on empirical methods in Natural Language Processing, Vol. 10, pp. 79-86, 2002
[3] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts", Proceeding of the 42nd Annual Meeting on Association for Computational Linguistics, ACL, 2004.
[4] M. Rushdi Saleh, M. T. Martin-Valdivia, A. Montejo-Raez, L. A. Urena-Lopez, "Experiments with SVM to classify opinions in different domains", Expert Systems with Applications 38, pp. 1138-1152, 2011.
[5] Rodrigo Moraes, Joao Francisco Valiati*, Wilson P. G. Neto, "Document Level Sentiment Classification: A Empirical Comparison between SVM and ANN", Elsevier 2012.
[6] Alexander Pak, Patrick Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", Proceedings of the Seventh conference on International Language Resources and Evaluation, pp. 1320-1326, 2010.
[7] Rui Xia, Chengqing Zong, Shoushan Li, "Ensemble of feature sets and classification Algorithms for Sentiment Classification", Information sciences 181, pp. 1138-1152, 2011
[8] A. Nisha Jebaseeli, E. Kirubakaran, "Neural Network Classification Algorithm with M-Learning reviews to improve the Classification Accuracy", IJCA, Vol. 71, June 2013.
[9] Guang Qiu, Bing Liu, Jiajun Bu and Chun Chen, "Opinion Word Expansion and Target Extraction through Double Propagation", Computational Linguistics, March 2011, Vol. 37, No. 1: 9.27.
[10] Suke Li, "Sentiment Classification using Subjective and Objective Views", IJCA, Vol. 80, No. 7, Oct 2013.
[11] X. Glorot, A. Bordes, Y. Bengio, "Domain Adaptation for Large Scale Sentiment Classification: A Deep

Learning Approach", Proceedings of the 28th International Conference on Machine Learning, 2011.
[12] Gang Li, Fei Liu, "A clustering based approach on Sentiment Analysis", IEEE 2010.
[13] Gang Li, Fei Liu, "Sentiment Analysis based on Clustering: a framework in improving accuracy and recognizing neutral opinions", Vol. 40, pp. 441-452, Springer 2013.
[14] M. Weigand, A. Balahur, A. Motoyo, B. Roth and D. Klakow, "A Survey on the role of Negation in Sentiment Analysis", Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, pages 60–68, Uppsala, July 2010.
[15] S. Chandrakala and C. Sindhu, "Opinion Mining ad Sentiment Classification: A survey," ICTAT Journal on Soft Computing, ISSN: 2229-6956, 2012.

## Author Profile

**Archana Sonagi** received the B.E degree in Information Technology Engineering from M.G.M's JNEC, B.A.M.U in 2012. She is now pursuing M.E. degree in Computer Engineering from P.E.S.'s Modern College of Engineering, Savitribai Phule Pune University, Pune.

**Deipali Gore** received the M.E degree in Computer Engineering from D.Y. Patil, Akurdi College under Savitribai Phule Pune University in 2006. Her current research area includes Information Retrieval and Web Mining. She is currently working as an Assistant Professor in P.E.S.'s Modern College of Engineering, Savitribai Phule Pune University, Pune.