

Full Length Research Paper

Sentiment analysis as a way of web optimization

Osama M. Rababah^{1*}, Ahmad K. Hwaitat², Dana A. Al Qudah¹

¹Department of Business Information Technology, the University of Jordan, Jordan.

²Department of Computer Science, the University of Jordan, Jordan.

Received 8 January, 2016; Accepted 14 April, 2016

Web optimization is the process of optimizing the web to increase visibility or rank of websites in search engines. Furthermore, this process is also viewed from multiple perspectives, from optimizing inter-server communication that offers the best responses to users' queries and provides targeted advertisements to users of a website. With this regard, the process of automatic classification and information extraction from users' comments, also known as Sentiment Analysis (SA) or opinion mining, becomes vital to offer users the best online experience, based on their preferences. There are numerous algorithms available for SA. Therefore before applying any algorithm for polarity detection, pre-processing on comments is carried out. This study analyzes how we can write an algorithm for performing SA, and how different types of processing that are applied to initial data such as stemming or eliminating stop words affect the performance of this algorithm. The results show that even when a small sample is used, sentiment analysis can be done with a high accuracy (over 70%) if appropriate natural language processing algorithms are applied. Having a method for guessing sentiments could enable us, to excerpt opinions from the internet and predict online customer's favorites, which could ascertain valuable for commercial or marketing research.

Key words: Sentiment analysis, natural language processing, python programming language, machine learning, web optimization.

INTRODUCTION

Sentiment analysis (SA) is the use of natural language processing, statistics, or machine learning approaches to extract, identify, or otherwise describe the sentiment content of a text unit. SA or opinion mining is the computational study of people's politics, attitudes, and emotions to elements such as topics, products, individuals, organizations, services, etc. SA can be valuable in several ways. For example, in marketing it helps in determining the success of an ad campaign or new product launch, determine the versions of a product

or service are accessible and even identify which demographics like or dislike particular features (Chenlo and Losada, 2014).

Sentiment analysis is not only applied to product reviews but can also be applied to stock markets (Hagenau et al., 2013), news articles (Xu et al., 2012) or political debates (Maks and Vossen, 2012). In political debates, for example, we could analyze trends, identify ideological bias, target advertising/messages, gauging reactions, etc. The election results can also be predicted

*Corresponding author. E-mail: o.Rababah@ju.edu.jo.

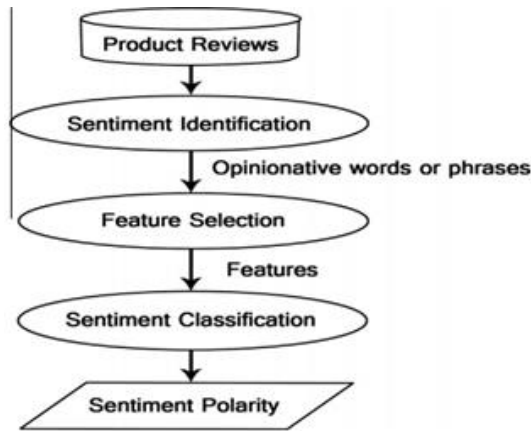


Figure 1. Sentiment analysis process on product reviews (Medhat et al., 2014).

from political posts. The social network sites and micro-blogging sites are considered a good source of information because people share and discuss their opinions about a certain topic freely. They are also used as data sources in the SA process.

According to Medhat et al. (2014), SA can be considered a classification process as illustrated in Figure 1. There are three main classification levels in SA: document-level, sentence-level, and aspect-level SA. Document-level SA aim to categorize an opinion document as expressing a positive or negative opinion or sentiment. It considers the entire record a basic information unit. Sentence-level SA seeks to establish opinion expressed in each sentence. The first step is to identify whether the sentence is subjective or objective. If the decision is personal, sentence-level SA will decide whether the sentence expresses positive or negative opinions. Aspect level is also known as phrase-level opinion mining. It performs fine-grained analysis and directly looks at the opinion. The goal of this level of analysis is to discover feelings on aspects of items (Selvam and Abirami, 2009).

There are numerous challenges in sentiment analysis. The first is an opinion word that is considered to be positive in one state might be regarded as negative in another state. The second challenge is that people do not always express opinions in the same way. Most traditional text processing depends on the fact that small differences between two pieces of text do not alter the meaning. People can be contradictory in their statements (Selvam and Abirami, 2009).

People are envisioned to develop an algorithm that can identify and classify opinion or sentiment as represented in an electronic text. There are many applications and enhancements on SA algorithms that were proposed in the last few years. The purposes of this paper are to analyze how an algorithm for accomplishing SA can be written and how its performance is affected by different

types of processing.

STATE OF THE ART

Abundant researches exist on SA of user belief data (Pang and Lee, 2008; Montoyo et al., 2012; Tsytsarau and Palpanas, 2012; Cambria et al., 2013; Feldman, 2013; Medhat et al., 2014; Serrano-Guerrero et al., 2015) which primarily judge the polarities of user reviews. In these studies, sentiment analysis is often conducted at one of the three levels: the document level, sentence level, or attribute level (Jeyapriya and Selvi, 2015).

Pang and Lee (2008) on their research focused on the applications and challenges in SA; they mentioned the techniques used to solve each problem in SA. Tsytsarau and Palpanas (2012) have presented a study that discussed the main topics of SA in details. For each topic, they have illustrated its definition, problems, and development and categorized the articles with the aid of tables and graphs. Medhat et al. (2014) classify a large number of recent articles according to the techniques used with brief details of the algorithms and their originating references.

The literature survey done indicates two types of techniques including machine learning and semantic orientation. These methods are shown in Figure 2.

Supervised learning approach contains two sets of documents training and test. To learn about the document, the training set is used by classifier. For validation test, set is used. For review classification, many techniques can be employed. Types of supervised learning methods include decision tree classifier, linear classifier, rule based classifier, and probabilistic classifier as shown in Figure 2.

In dictionary-based approach, a small set of sentiment words which are identified as seed words are composed manually with their known positive or negative orientations. After that, this set is grown by searching their synonyms and antonyms in an online dictionary. The new words are added to the on hand seed list. Then, the next iteration takes place. The iteration should be closed when no new words are found. Finally, manual inspection set is used to clean up the list (Jeyapriya and Selvi, 2015).

SA can be analyzed using various initial data corpuses and methods of classification. One type of approach is to divide the data into a given number of categories. An example of this is Turney (2002), where the original data was classified into two categories (polarity analysis) based on the data present in online reviews. Another type of sentiment analysis involves groups that are related to each other, and the degree of positivity or negativity of an individual comment makes it belong to either group. An example of this type of approach is the classification of movie reviews on a scale of 0 to 4 stars (Pang and Lee, 2004). This paper aims to examine how an algorithm for accomplishing SA can be written and how its performance

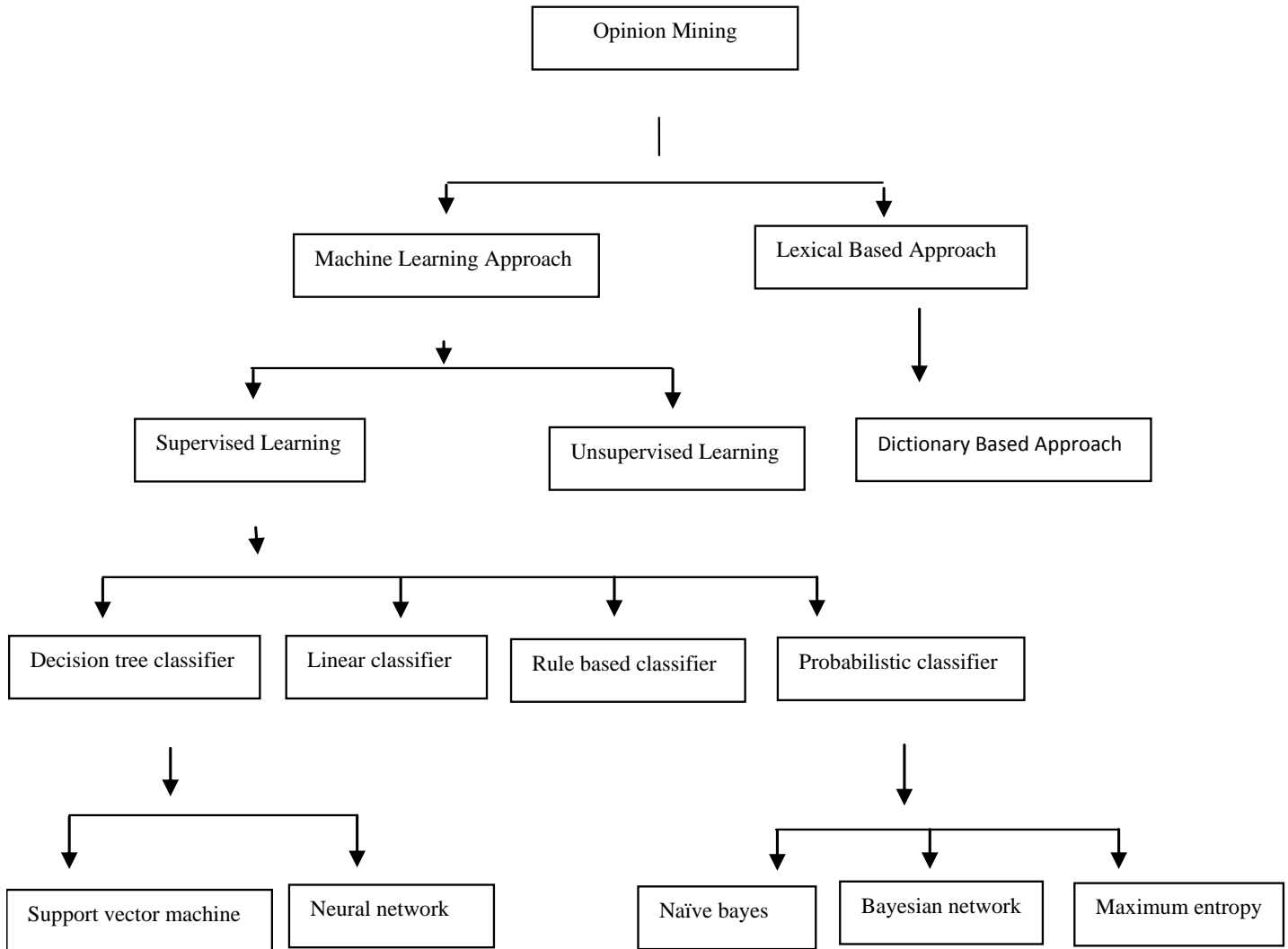


Figure 2. Opinion mining techniques.

is affected by different types of handling.

All kinds of sentiment analysis have natural language processing as a primary component. Some means in which natural language can be processed for usage in this kind of algorithms can be seen subsequently.

Means of data processing

Basic data processing

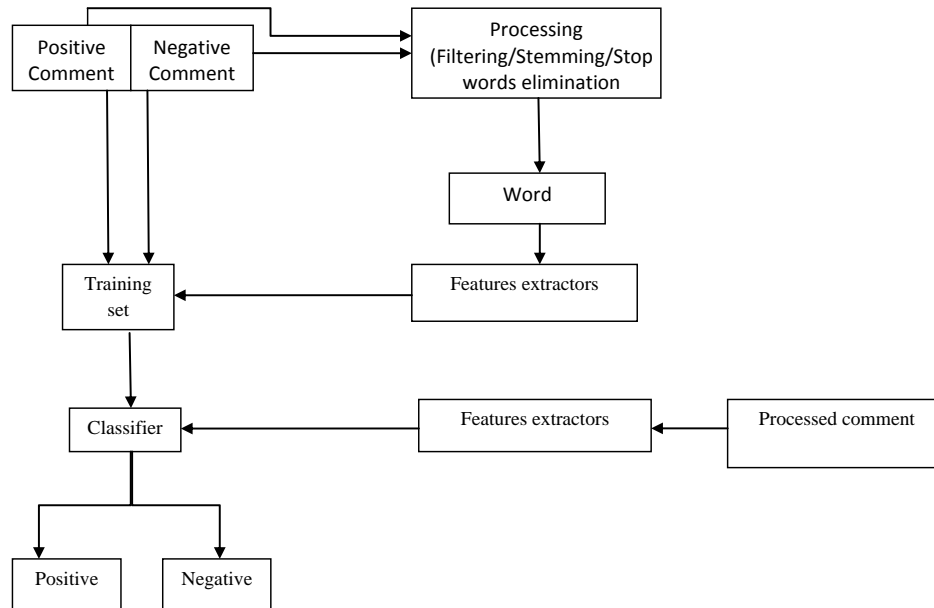
The primary data processing consists of eliminating case sensitivity for all the words present in the dataset (training and test set). This is done to ensure that the positions of a word within the sentences in which it appears do not influence its statistical relevance, so it is capitalized, and non-capitalized forms are counted together (Aggarwal, 2015).

Stemming

Stemming is a procedure that is used to group together words that are similar in meaning and written form (different verbal forms, adjectives, and nouns that are derived from a single term). After being reduced to the stem form (the root of all derivations), all appearances of the words are counted together to be used later in the statistical analysis. After deriving the meaning of the texts becomes more easily linked to the phrase used, and the syntactic-semantic gap becomes easier to handle (Aggarwal, 2015).

Stop word elimination

Another measure taken to eliminate or at least reduce the syntactic-semantic gap is the removal of stop words from



Scheme 1. Sentiment analysis schema.

the text. Words that are very frequent in any text and carry little meaning by themselves, such as articles, some verbs, and other similar structures, constitute stop words.

The stop words are eliminated so that words that hold more meaning are given greater importance when the analysis and classification are made (Aggarwal, 2015).

ALGORITHM

Description

The tagged and classified data consisting of comments that were previously identified as positive or negative is processed through various methods (including stemming and stop word elimination) during the processing phase. Then the words that hold the most statistical significance for the meaning and polarity of a text are identified.

The feature extraction phase composed of obtaining a feature list, consisting of indicators of the presence or absence of any of the words that were previously identified within each of the texts in the training set.

The feature lists are then passed to a machine learning algorithm that gets trained with the already classified training data and classifies the comments from the test sets into positive or negative ones, based on its information. Scheme 1 shows how a sentiment analysis algorithm works and how the data is processed, for both the training set and the test set (Pang and Lee, 2015).

METHODS

For this study, 200 tagged positive and 200 negative reviews from

the corpus consisting of movie reviews introduced by Pang and Lee (2004) were analyzed. The data was divided into an 80% training set and a 20% test set, against which the algorithm was tested.

The Naïve Bayes classifier from the nltk package (Bird et al. 2009) in Python was used as a classifier. The Naive Bayes classifier is the simplest and most commonly used classifier. Naive Bayes classification model computes the posterior probability of a class, based on the distribution of the words in the document. The model works with the BOWs feature extraction which ignores the position of the word in the document. It uses Bayes theorem to predict the probability that a given feature set belongs to a particular label.

The Stop Word list that was used was also the one present in the nltk package and the stemmer used was the snowball stemmer found in the documentation of the same package. The processing part of the algorithm was different for each of the five instances of the algorithm that were run:

NoProc: The words in the original data were not modified in any way before analysis;

MinProc: Capitalization was eliminated;

Stemming: The words in the original data were reduced to their stems through the Snowball Stemmer;

StopWordEliminated: The relevant words were selected from the original data by eliminating the stop words from the nltk stop word list;

StopWordEliminated and Stemming: The proper words were chosen from the original data by removing the stop words that were present and then were reduced to their stem forms.

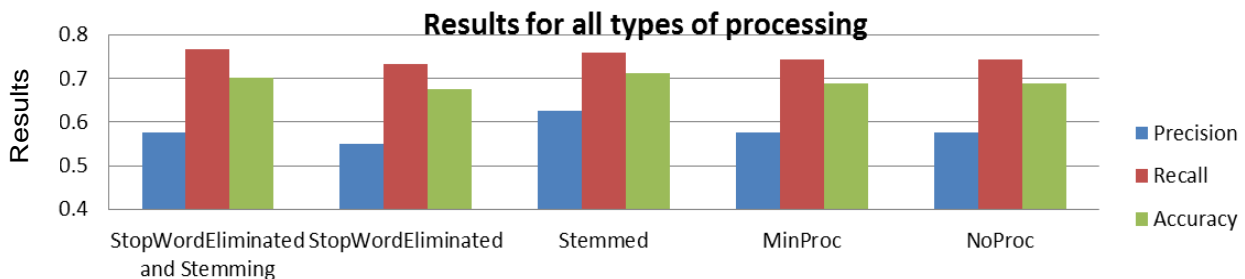
The results obtained through all of the five instances were analyzed using the relevant measurements that are described in running the code and interpretation of the results.

Relevant measurements

When classifications and data retrieval are concerned, the algorithms' performances need to be analyzed according to three

Table 1. Relevant measures for the classification.

Parameter	Precision	Recall	Accuracy
StopWordEliminatedand Stemming	0.58	0.77	0.70
StopWordEliminated	0.55	0.73	0.68
Stemmed	0.63	0.76	0.71
MinProc	0.58	0.74	0.69
NoProc	0.58	0.74	0.69

**Figure 3.** Graphical representations of results.

specific measures that will be explained.

Recall

Recall is the true positivity rate of the classification, also referred to as sensitivity. Recall can be calculated as:

$$\frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Precision

The precision is calculated as the positive prediction value, having the formula:

$$\frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Accuracy

Accuracy is the rate of which items have been correctly classified and/or retrieved. It can be calculated as:

$$\frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{False Negatives} + \text{True Negatives} + \text{False Positives}}$$

The values that were obtained for each of these indicators are shown and discussed in the results of this paper.

RUNNING THE CODE AND INTERPRETATION OF THE RESULTS

The project described in this article was developed using the python programming language and the NLTK package.

For running the code, the NLTK package and its corpuses and packages (stop word list, stemmer, etc.) need to be installed. The feature extractor was used to process the data and divide the resulting list of feature sets into a training set and a test set. The training set is used to train a new "naive Bayes" classifier. In case another corpus needs to be used for training or testing, the tagged comments need to be added to the pos or neg folders (depending on their polarity) present in the project folder, each of the comments needs to be written in a new file, whose name is not important.

The results obtained after the run of the algorithm are automatically saved into log files with names that include the type of processing that was used on the data. The log files consist of many lines that are equal to the number of test cases that were provided, each line composed of the absolute polarity of the comment followed by the polarity that was assigned to it by checking it with the algorithm. An example of a line within a log file is: p,n

This means that the comment that was tested was tagged as being positive in the corpus, but the algorithm has wrongly classified it as being a negative one. The data in these logs was later processed and analyzed for the particular classification measurements to be calculated: recall, precision, and accuracy.

RESULTS AND DISCUSSION

The results that were obtained when the algorithm is run in all five instances are shown in Table 1. A graphical representation of these results is shown in Figure 3. It can be observed from both Table 1 and Figure 1 that the Recall improves as the text is processed more (the best recall is obtained when both the stop words are eliminated and stemming is used). The other two indicators (precision and accuracy) have the maximum values when stemming only is used as a processing algorithm. Figure 4 shows how the accuracy of the classification is affected by the different processing methods that were used for

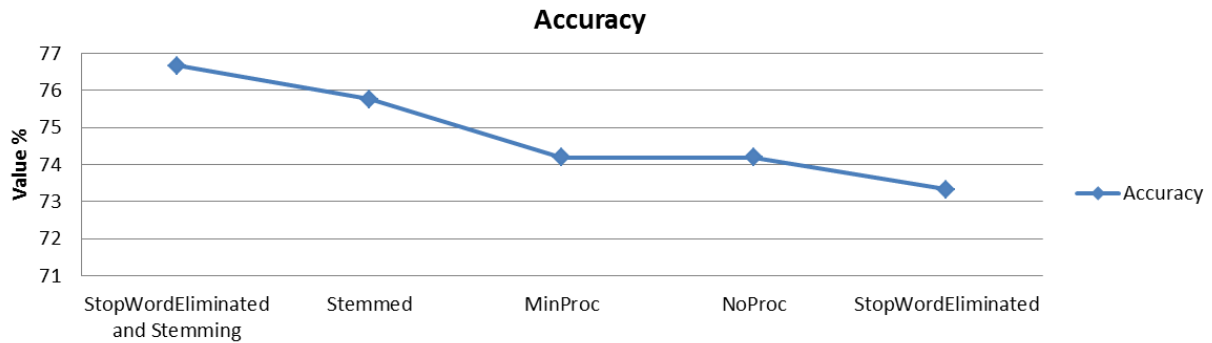


Figure 4. Accuracy dependence on the processing method used.

the data.

It can be observed that, for the type of data that was used for this trial classification, the elimination of stop words is not a viable technique by itself, as it decreases the accuracy of the classification, but, when used alongside with stemming, it gives the best results. A positive or negative sentiment word may have their conflicting meaning in a particular domain, so it is hard to predict by its keyword meaning. An interrogative sentence may not have neither positive nor negative sentiment but the key word used in the opinion may be positive or negative.

Rare sentences in the form of jocks may disrupt the meaning of the entire sentences; such type of sentence needs a powerful attention toward the keywords and sentences. These funny sentences not only violet the term of a specific sentence, but also destroy the value of the complete document.

Sometimes, sentiments do not use any sentiment words like good, better, best, worst, bad, etc., but the sentences may have its positive or negative feedback about the product, services, and policies. Conditional sentences are also an issue in sentiment mining. Conditional sentences are also creating the same problem like interrogative sentences.

Conclusions

The present paper shows the connection between five different methods (No processing, Minimal processing - elimination of capitalization, Stemming, Elimination of Stop Words and Stemming used after the Elimination of Stop Words) of natural language processing and their effect on the efficiency of a sentiment analysis algorithm which provides two categories of entries: positive or negative.

Considering that the size of the sample with which the machine learning algorithm was rather small, and the mean dimension of an entry was rather large and used varied language (the entries consisted of positive and negative movie reviews), it was expected that the

categorization of the test data would not be perfect, at around 60%. On the other hand, the accuracy and recall had values that were higher than 70%.

The best results in recall were obtained when the data was processed by both stemming and elimination of stop words while the best accuracy was observed when only stemming was used for processing. Minimal processing (reducing everything to lower cases) does not influence the results in any way in this case, as the original data was already pre-processed within the samples. These findings show that, even when a small sample is used, sentiment analysis can be done with a high accuracy (over 70%) if appropriate natural language processing algorithms are applied.

In this project, only individual words were considered for statistical analysis and determining the meaning and tone of the comments. As a next step, bigrams and trigrams (groups of two and respectively three words used in a given order) could also be utilized for the data classification.

Conflict of Interests

The authors have not declared any conflict of interests.

REFERENCES

- Aggarwal CC (2015). Data mining: The textbook. Springer.
- Bird S, Edward L, Ewan K (2009). Natural Language Processing with Python. O'Reilly Media Inc. Available at: <http://www.nltk.org>. Retrieved 03, 2015.
- Cambria E, Schuller B, Xia Y, Havasi C (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intell. Syst.* 1(2):15-21.
- Chenlo JM, Losada DE (2014). An empirical study of sentence features for subjectivity and polarity classification. *Inf. Sci.* 280:275-288.
- Feldman R (2013). Techniques and applications for sentiment analysis. *Commun. ACM* 56(4):82-89.
- Hagenau M, Liebmann M, Neumann D (2013). Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Syst.* 55(3):685-697.
- Jeyapriya A, Selvi K (2015). Extracting aspects and mining opinions in product reviews using supervised learning algorithm. In: *Electronics and Communication Systems (ICECS), 2015 2nd International Conference on IEEE*. pp. 548-552.

- Maks I, Vossen P (2012). A lexicon model for deep sentiment analysis and opinion mining applications. *Decis. Support Syst.* 53(4):680-688.
- Medhat W, Hassan A, Korashy H (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Eng. J.* 5(4):1093-1113.
- Montoyo A, MartiNez-Barco P, Balahur A (2012). Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments. *Decis. Support Syst.* 53(4):675-679.
- Pang B, Lee L (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics. P 271.
- Pang B, Lee L (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval* 2(1-2):1-135.
- Pang B, Lee L (2015). Polarity dataset v2.0 (3.0Mb). Available at: <http://www.cs.cornell.edu/people/pabo/movie-review-data/> Retrieved December 03, 2015.
- Selvam B, Abirami S (2009). A Survey on Opinion Mining Framework. *Int. J. Adv. Res. Comp. Commun. Eng.* 2(9):3544-3549.
- Serrano-Guerrero J, Olivás JA, Romero FP, Herrera-Viedma E (2015). Sentiment analysis: A review and comparative analysis of web services. *Inf. Sci.* 311:18-38.
- Tsytarau M, Palpanas T (2012). Survey on mining subjective data on the web. *Data Mining Knowledge Discovery* 24(3):478-514.
- Turney PD (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: *Proceedings of the 40th annual meeting on association for computational linguistics*. Assoc. Computat. Linguist. pp. 417-424.
- Xu T, Peng Q, Cheng Y (2012). Identifying the semantic orientation of terms using S-HAL for sentiment analysis. *Knowledge-Based Syst.* 35:279-289.