DOI: http://dx.doi.org/10.26483/ijarcs.v8i7.4095

ISSN No. 0976-5697

Volume 8, No. 7, July – August 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

VIABLE MODERN APPROACHES FOR SENTIMENT ANALYSIS: A SURVEY

A Srinivas
Research Scholar
Rayalaseema University
Kurnool, Andhra Pradesh, India

MHanumanthappa
Professor, Dept. of Computer Science & Applications
Bangalore University, Jnanabharathi Campus
Bangalore, India

Abstract: Sentiment analysis is a process of extracting, identifying and categorizing a writer's emotion, expressed in the form of text, by implying a computational method. This paper presents a study of various modern approaches for sentiment analysis along with the hurdles and possible solutions present in these approaches. Further, the study is concentrated on two main categories, machine learning and lexicon analysis, for sentiment analysis. Even though, the various methods falling under these two main approaches are elaborated and illustrated, the supervised learning method of machine learning is more concentrated in the article. This paper also describes a generalized sentiment analysis method that can be incorporated with any of the existing analytical algorithms for sentiment analysis as well as any mundane text analysis.

Keywords: Sentiment analysis; machine learning; lexicon analysis; supervised learning; text analysis.

I. INTRODUCTION

Sentiment analysis is one of the most worked subfield of Natural Language Processing (NLP) and has seen a considerable amount of research done during past decade. It is a process of extracting, identifying and categorizing an emotion from text data. Emotion artificial intelligence or opinion mining are some alternative terms used for sentiment analysis. Sentiment analysis is very important from business perspective of social media or any product based marketing. The social media giants like twitter, facebook, pinterest etc. are all heavily depending on sentiment analysis to make most of their applications. Many algorithms are being used for sentiment analysis, fall under either of the two well established computing approaches: Machine Learning and Lexicon analysis.

Machine learning approach to the sentiment analysis can be done in two ways. The first one, supervised learning in its simplest form, is a type where to produce a predicted outcome training data sets are used on a set of known inputs. A well-known algorithm is used to carry out this process. The output is already known in supervised learning. It is one of the mostly implemented machine learning methods in modern industries. Decision trees, linear and rule based classifiers fall in this category of approach.

The second one is called as unsupervised learning, where there are no training data sets and outputs are known. That is why it is a more complex method and is being used in far smaller number of applications so far. In unsupervised learning, an Artificial Intelligence (AI) agent goes into the problem without having any prior knowledge about the problem.

The other approach to sentiment analysis apart from machine learning is Lexicon-analysis. Dictionary-based approach is one of the methods under lexicon-analysis. Few words which have maximum influence on the output of the content in term of sentiment score are collected and their possible synonyms and antonym are searched in well-known dictionary repositories like thesaurus or WordNet etc. The word seed will be added with these newly found synonyms and are used every time the word is used in searching.

In Corpus-based approach of Lexicon-analysis, opinion words with context specific orientation are searched depending on syntactic patterns or patterns that occur together along with a seed list of opinion words to find other opinion words in a large corpus [1]. Here a corpus stands for a collection of written relative text. The figure 1 shown below depicts various algorithms falling under these two approaches.

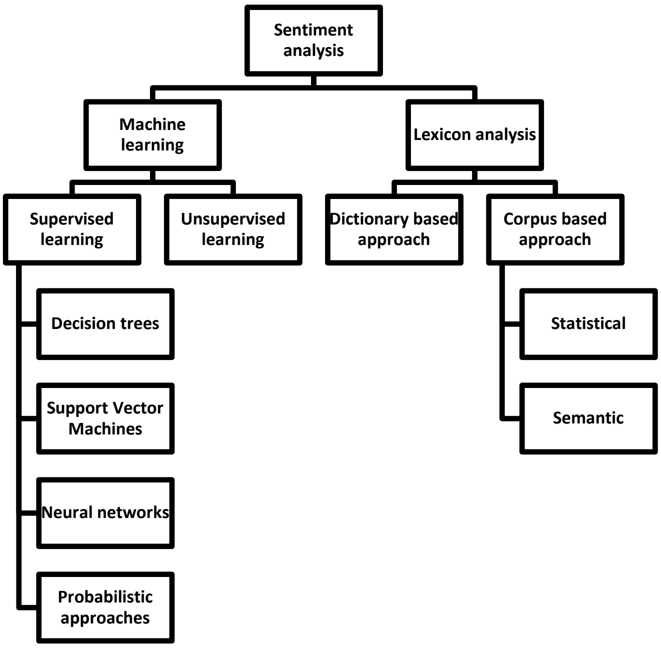

Figure 1. Approaches to sentiment analysis.

II. RELATED WORK ON APPROACHES TO SENTIMENT ANALYSIS

In early 2000s notable works from Prof. Jason Eisner of JHU and Prof. Lillian Lee of Cornell sparked the word "Sentiment Analysis". Later in 2002 a preceding of Conference on Empirical Methods on Natural Language Processing presented by Prof. Bo Pang and his colleagues titled "Thumbs up? Sentiment classification using machine learning techniques" initiated worldwide research work on sentiment analysis [2]. Later in 2004 first effort to bringing together various approaches – machine learning, lexical,

© 2015-19, IJARCS All Rights Reserved

knowledge based etc. – were made inassociation for the Advancement of Artificial Intelligence (AAAI) spring symposium.

Later efforts turned to a more polar view of sentiment, from positive to negative, such as work by Peter D Turney and Bo Pangwho applied different methods for detecting the polarity of product reviews and movie reviews respectively. This work is at the document level. One can also classify a document's polarity on a multi-way scale, which was attempted by Bo Pang and Benjamin Snyder among others: Bo Pang and Lillian Lee expanded the basic task of classifying a movie review as either positive or negative to predict star ratings on either a 3- or a 4-star scale, while Snyder performed an in-depth analysis of restaurant reviews.

Even though in many of the statistical classification procedures, the neutral class is ignored with the assumption that neutral texts are closely bound with the binary classifiers, several researchers suggest that, as in every polarity problem, three categories (namely a positive, a negative and a neutral) must be identified. Moreover, it can be proven that specific classifiers such as the Max Entropy [3] and the SVMs [4]can benefit from the introduction of a neutral class and improve the overall accuracy of the classification.

There are in principle two ways for operating with a neutral class. Either, the algorithm proceeds by first identifying the neutral language, filtering it out and then assessing the rest in terms of positive and negative sentiments, or it builds a three-way classification in one step. This second approach often involves estimating a probability distribution over all categories (ex. Naive Bayes classifiers). Whether and how to use a neutral class depends on the nature of the data: if the data is clearly clustered into neutral, negative and positive language, it makes sense to filter the neutral language out and focus on the polarity between positive and negative sentiments. Recently due to the rapid growth of social media, sentiment analysis has a huge impact on social media marketing. Many universities and private organizations are contributing a lot of research work in the field of sentiment analysis.

III. GENERALIZED METHOD FOR SENTIMENT ANALYSIS

Even though a generalized method is difficult to present for any set of problem, sentiment analysis can be exempted from such declaration. Due to the similarity between the types of data used in analysis, it is possible to propose a generalized method for sentiment analysis.

Following are the different phases in the proposed generalized method for sentiment analysis:

A. Data acquisition

Dataset for analysis would comprise of data in text format including social media data, user reviews from various websites. Most of these dataset corpuses have been provided by various universities and research scholars from all over the globe (For ex: Twitter sentiment corpus by Niek Sanders).

Representation of dataset may vary due to the combination of techniques used in further processing of the data. JSON (JavaScript Object Notation) is an interesting data interchange format that can be used for various type of data representation.

B. Text preparation

Preparing text for the analysis feed is a little tricky task due to the presence of sarcasm, metaphor like contents in datasets. Text preparation is a two-step process including POS tagging and an optional step of filtering for contents which has no effect on the sentiment analysis result.

In corpus linguistics, part-of-speech tagging (POST or POS tagging), also called grammatical tagging or word-category disambiguation, is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context—i.e. relationship with adjacent and related words in a phrase, sentence, or paragraph.

A simplified form of this is used in the identification of words as nouns, verbs, adjectives, adverbs, etc. Filtering a text for useful content involves tokenization and normalization. Tokenization is the name given to the process of chopping up sentences into smaller pieces (words or tokens). The segmentation into tokens can be done with decision trees or customized tokenizers like openNLPtokenizer, Stanford tokenizer etc.

Text filtering is also carried out during this phase. For example, few words like ‘craaaazy’ need to be preprocessed for reducing it to ‘crazy’ to make sentiment analysis effective.Stop word removal is another part of text preparation. This step is removal of adverbs, articles, prepositions from the text which do not contribute to the sentiment analysis. Stemming can also be done on the text to reduce a word to its root form. For example, talking and talker have common root word talk. So, the stemmed form of both these word is talk.

C. Sentiment detection

Detecting a sentiment and categorizing the same is the most challenging part of this process. This is the point where either machine learning or a lexicon-analyzing technique has to be used. Large dictionary corpuses like Thesaurus or WordNet are very useful in normalization of words. This dictionary based approach will be used to detect sentiment from a dataset feed.

D. Sentiment classification

Classification of sentiment is done based on sentiment scores. The polarity of the content plays very important role in deciding whether it is neutral, positive or negative. This polarity is decided based on the score. For example, on a scale between 1 to 10, if a content is given with a value of 5 then it can be considered as neutral, if the value is less than 5 it can be a negative sentiment or if the value is more than 5, it can be a positive one.

Due to the importance of sentiment analysis in social media marketing, a decent number of tools, web applications, APIs have been framed from various service providers to handle sentiment analysis.

E. Result analysis

The last stage of the sentiment analysis is to analyze the result for its accuracy and present the proof for the same using analytical tools. There is no guaranteed tool or method to present sentiment analysis with 100% accuracy. Several reasons can be found for this. One, due to the inclusion of metaphors, sarcasm in the data being fed to the analysis, it is difficult to make a computer system understand those type of data. Two, choosing what defines polarity has always been difficult. This part is crucial because polarity is the one that decides final sentiment shown in the data.

Result analysis can be carried out in many ways. Plotting a graph is one way and bar charts have been other better options for the purpose. Result analysis can be used to verify the effectiveness of the methods being used. It shows the reliability of the method used for analysis.

Following figure 2 shows all of the above mentioned phases with illustration:

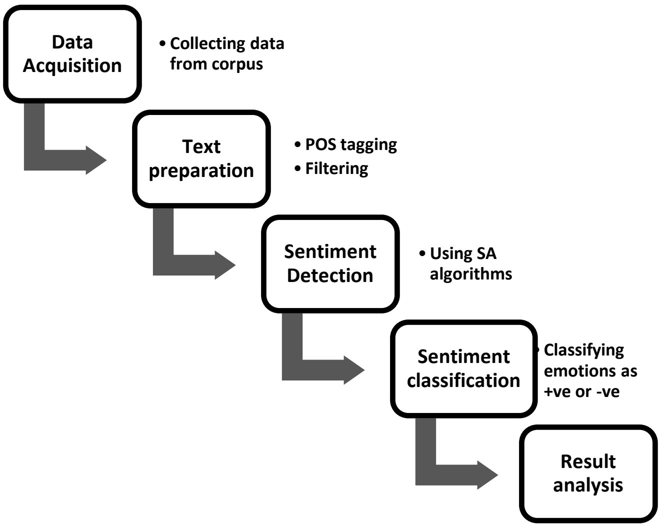

Figure 2. Generalized method phases.

IV. APPROACHES TO SENTIMENT ANALYSIS

The first bands of methods for sentiment analysis are from supervised learning of machine learning. Compared to its counterpart unsupervised learning, supervised learning methods are easy to implement because of known labeledinputs and predictable outputs. Second in line are lexicon-based approaches. Following section has some details about machine learning as well as lexicon-based approach.

A. *Supervised learning*

1) Decision trees

Decision trees (DT) are supervised learning methods used for classification and regression. A tree representation is used in the process where a prediction about an item's target value(represented by leaves) is made by observing an item (represented by branches) [5]. Due to the predictive nature of the method it is very useful in statistics and data mining. Decision trees are in existence for a long period now. A new approach to decision trees called as a semi-supervised decision trees are being used as a substitute to them. It is a type of technique that can take both labelled and un-labelled data as input.

DT is a classic technique for sentiment analysis. Texts can be classified using some defined features or sentiments. The challenge in using a DT for sentiment analysis is to define a parameter using which classification can be done.

Decision trees are easiest implementation methods in machine learning because,

- Interpretation is very easy due to tree structure which can be easily visualized.
- Little or no data normalization. DTs do not support any missing or incomplete data values. So, there is no required pre-processing or normalization of data before plotting decision trees.
- The cost of using the number of representational data points in a tree formation is logarithmic. Due to this, implementation of DT usually has no burden on the application using it.

Apart from being easiest DT have some disadvantages to be addressed,

- Over-fitting is one of the most common problem faced with a DT implementation. One of the purposes of DT is to generalize data but one may create much complex tree which does not serve this purpose. Such conditions are called as over-fitting. Setting the maximum depth value for a tree can be used to avoid this problem.
- Another problem with DT is instability. A small change or variation in value of one of the node may lead to generation of a whole new tree. Static inputs and trees within an ensemble can be used to address this issue.

2) Support vector machines

Support vector machines (SVM) is a classififcation technique in machine learning that can act on a set of labelled data to classify it using defined parameters. SVM uses something called as a hyperplane or decision plane to classify these values [6]. A hyperplane is the deciding factor to all the classification done using SVM. There is a possibility of defining more than one such hyperplanes for a set of labelled data. The classification is represented on n-dimensional graph where each dimension is a defined feature of those set of objects.

An optimal hyperplane would be the one that has highest marginal gap between itself and the nearest plots of the graph.

The following figure 3 represents a graph used to plot different comments using their sentiment scores. These sentiment scores are the parameters used to classify comments. The line dividing two groups is the hyperplane.

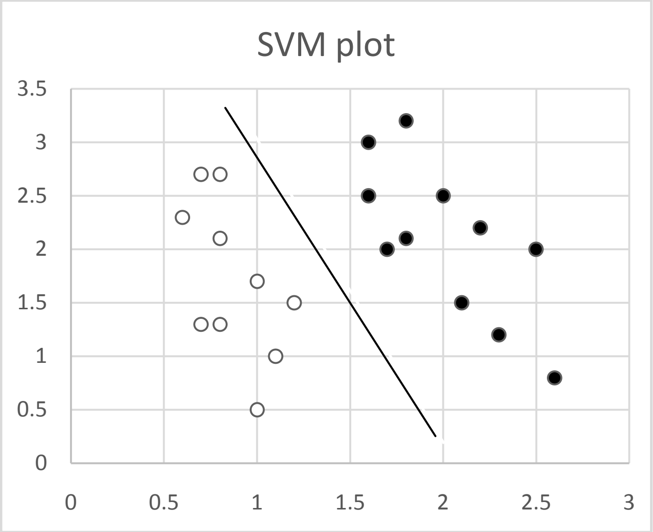

Figure 3. SVM plot illustrating hyperplane.

Figure 4. Following are some advantages of SVM implementation,

Figure 5. SVM is very effective for high dimensional spaces. Also it is very effective in cases where number of dimensions are greater than the number of samples. More the margin between the hyperplane and the nearest plotted data, more effective will be the result of classification.

Figure 6. The support vectors (the subset of training points in the decision function) can make very effective use of memory. So implementation of SVM is very cost effective.

Figure 7. SVM still possesses some flaws which are mentioned below,

Figure 8. SVM doesn't perform well when data set is very large. This is because for a huge data set the training data takes a lot of time.

Figure 9. SVM cannot handle noise very effectively. Data has to be very well defined. Any overlapping objects in the data could easily produce undesired results.

3) *Neural networks*

Artificial neural networks (ANN) or neural networks are computing systems inspired by the biological neural networks. The sytem built on the basis of a neural network is designed to improve its performance progressively. Ideally, ANN is designed to learn from examples. It can extract patterns and detect trends that are too complex to be noticed by either a human being or a computer. Due to this ability it is largely used in the field of pattern recognition and deep learning [7].

Sentiment analysis can be done using neural networks in the same way it is used for pattern recognition. Since there is no need to devise an algorithm, it is easy to implement neural network provided that the methodology is clearly stated.

It is difficult to compare ANN with a conventional computation because unlike conventional computing methods, ANN's approach has no fixed, well-defined algorithms. But its implementation is considered as one of the most interesting approach for sentiment analysis. Following are some advantages of ANN related to sentiment analysis,

- Easy to use and understand since it has non-linearity in its structure.
- Few of the algorithms of neural networks like Back Propogation algorithm are widely used because of their learning ability.
- If there is a strong hardware support ANN could be prove itself as best solution to most of the problems.

Now, neural networks may be considered as one of the newest approach to a computational problem but it comes with several disadvantages too,

- The dataset needs to be trained with a huge number of examples before the processing. More the training more accurate the result. This may lead to consumption of time.
- A neural network is a black box design. So nothing much can be gleaned off from this method.

4) *Probabilistic approaches*

Probabilistic classifiers are developed by assuming generative models which are product distributions over the original attribute space (as in naive Bayes) or more involved spaces (as in general Bayesian networks) [8]. There are three well-known methods in this category. Naïve Bayes, Bayesian networks and maximum entropy.

All probabilistic techniques take an input and predict a probability distribution instead of outputting a single class.The performances of techniques under probabilistic approach depend upon the size of training data.Overfit problem is very common if a proper method is not chosen.

The Naïve Bayes classifier is the simplest and most commonly used classifier in probabilistic approach. Naïve Bayes (NB) classification model computes the posterior probability of a class, based on the distribution of the words in the document. Another method in this category is Bayesian Networks (BN).The main assumption of the NB classifier is the independence of the features. The other extreme assumption is to assume that all the features are fully dependent. This leads to the Bayesian Network model which is a directed acyclic graph whose nodes represent random variables, and edges represent conditional dependencies. BN is considered a complete model for the variables and their relationships. Therefore, complete joint probability distributions (JPD) over all the variables are specified for a model. In Text mining, the computation complexity of BN is very expensive. That is why, it is not frequently used.

B. *Unsupervised learning*

Unsupervised learning is where there is only input data but no corresponding output or any predictions of output [9]. Unsupervised algorithms are left on their own to learn about input data. There are two broad classifications of unsupervised learning problems: clustering and association.

Clustering problems have huge application in the field of computer science. These are the problems where there is a possibility of detecting inherent groupings in the data. Many standard algorithms like k-means, hierarchical clustering is being used under this problem category.

Association problems are discovering rules or constraints that can describe large portions of data. This type of problem solving is very useful in social media marketing. Sentiment analysis can be carried out using such way of problem solving.

Unsupervised learning technique is least being used in sentiment analysis due to the difficulty in implementation. Due to the absence of a fixed structure to the process it is difficult to formulate algorithms for problems. Moreover, there is no point where one can predict the algorithm to stop. This is because of absence of predicting nature of the method. It is difficult to frame an output for the problem.

C. *Dictionary based approach*

This is a classic lexicon-analysis approach recommended by most of the sentiment analysis tools for POS tagging.WordNet is one large dictionary which has different synonymand antonym of words. In a dictionary based approach, a small set of sentiment words (seeds) with known positive or negative orientations is first collected manually [10]. An algorithm then grows this set by searching in the WordNet or another online dictionary for their synonyms and antonyms. The latterly found words are added to the seed list. The next iteration begins. The iterative process ends when no more new words can be found. After the process completes, a manual inspection step is used to clean up the list.

The approach is very simple. But its simplicity has many drawbacks. Firstly, metaphors don't have definitions stored in dictionaries. Due to the use of metaphors a dictionary may not be of any use to find synonyms. Secondly, it comes with a lot of manual work before and after the approach. Validating the seed list after one complete iteration is a difficult task.

D. *Corpus based approach*

There is a thin line difference between a corpus based approach and dictionary based approach. Instead of looking for a word or its synonym in a general dictionary, a corpus based approach would consider well defined corpora of sentiment words which don't require any validation [11]. This yields better accuracy in sentiment analysis.

Using the corpus-based approach alone is not as effective as the dictionary-based approach because it is hard to prepare a huge corpus to cover all English words, but this approach has a major advantage that can help to find domain and context specific opinion words and their orientations using a domain corpus. The corpus-based approach is performed using statistical approach or semantic approach as illustrated in the following subsections:

1) Statistical approach

Finding co-occurrence patterns or seed opinion words can be done using statistical techniques. This could be done by deriving posterior polarities using the co-occurrence of adjectives in a corpus [11].

It is possible to use the entire set of indexed documents on the web as the corpus for the dictionary construction. This overcomes the problem of the unavailability of some words if the used corpus is not large enough.The polarity of a word can be identified by studying the occurrence frequency of the word in a large annotated corpus of texts. If the word occurs more frequently among positive texts, then its polarity is positive. If it occurs more frequently among negative texts, then its polarity is negative. If it has equal frequencies, then it is a neutral word. But this may not be an ideal approach for SA since the outcome is dependent on how well the corpus is built.

2) Semantic approach

The Semantic approach gives sentiment values directly and relies on different principles for computing the similarity between words. This principle gives similar sentiment values to semantically close words. WordNet for example provides different kinds of semantic relationships between words used to calculate sentiment polarities. WordNet could be used too for obtaining a list of sentiment words by iteratively expanding the initial set with synonyms and antonyms and then determining the sentiment polarity for an unknown word by the relative count of positive and negative synonyms of this word.

The Semantic approach is used in many applications to build a lexicon model for the description of verbs, nouns and adjectives to be used in SA [12]. A model described in the reference [12] detailed subjectivity relations among the actors in a sentence expressing separate attitudes for each actor. These subjectivity relations are labelled with information concerning both the identity of the attitude holder and the orientation (positive vs. negative) of the attitude. Their model included a categorization into semantic categories relevant to SA. It provided means for the identification of the attitude holder, the polarity of the attitude and also the description of the emotions and sentiments of the different actors involved in the text. They used Dutch WordNet in their work. Their results showed that the speaker's subjectivity and sometimes the actor's subjectivity can be reliably identified.

V. DECISION CRITERIA FOR OPTING SENTIMENT ANALYSIS METHOD

Choosing a method for sentiment analysis would be a challenge. The general method gives a roadmap of the whole process but opting for the core sentiment analysis method will be kept open. Study suggests that the resources for sentiment analysis are reviews on products, places, events, etc. Nowadays, people are tending to express their emotion in complex ways by adding metaphors, sarcasm, slang words and ideograms (Ex: emoticons).

The approach has to be chosen manually since there is no method for an AI or computer to decide. Few parameters can be considered beforegoing for any of these approaches.

Factors like type of input data being used, size of the data, language used, source of the data etc. play a vital role in sentiment analysis. Different data types can be identified. It could be reviews, blogs, news, social media posts etc. Reviews are easy ones to analyze among the rest because the intension of a reviewer is to share his/her experience in terms of a sentiment. Blogs, news, posts are not only intended to express sentiment but also to share information. Thus giving more challenge to the process of sentiment extraction.

When it comes to size of the data, social media has a lion's share in providing the input for the analysis. Also social media is the one that has maximum benefit out of sentiment analysis in the form of social media marketing.

VI. CONCLUSION

This survey paper presents an overview of various modern algorithms used for sentiment analysis. Advantages and disadvantages of most of the algorithms have been discussed. The supervised learning approach of machine learning method is elaborated due to the recent popularity gain of these algorithms. Along with the survey a general methodology for sentiment analysis has been added in the study. The purpose of this general outline is to give a clear idea regarding sentiment analysis process and the necessary phases to be carried in the process.

VII. REFERENCES

[1] Douglas Rice, Christopher Zorn, "Corpus-based dictionaries for sentiment analysis of specialized vocabularies", New Directions in Analyzing Text as Data Workshop, London, ver 0.1, September 2013.

[2] Bo Pang, Lillian Lee, Shivakumar Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques", Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 79–86, 2002.

[3] Vryniotis, Vasilis, "The importance of Neutral Class in Sentiment Analysis", Machine Learning and Statistics, September, 2013.

[4] Koppel, Moshe, Schler, Jonathan, "The Importance of Neutral Examples for Learning Sentiment",Computational Intelligence 22. pp. 100–109, 2006.

[5] J. R. Quilan, "Induction of decision trees", Machine learning, pp. 81-106, 1986.

[6] M. K. S. Varma, N. K. K. Rao, K. K. Raju,G. P. S. Varma, "Pixel-Based Classification Using Support Vector Machine Classifier", Preceedings of the IEEE 6th International Conference on Advanced Computing (IACC), August, 2016.

[7] S. Samarasinghe, “Neural networks for applied sciences and engineering: from fundamentals to complex pattern recognition”, Aurebach Publications, April, 2016.

[8] Aggarwal Charu C, Zhai Cheng Xiang, "Mining Text Data", Springer New York Dordrecht Heidelberg London, Springer Science, LLC’12, 2012.

[9] Fu Xianghua, Liu Guo, Guo Yanyan, Wang Zhiqiang, "Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon", Knowledge Based Systems,pp. 186-195, 2013.

[10] Hu Minging, Liu Bing, "Mining and summarizing customer reviews", Proceedings of ACM SIGKDD international conference on Knowledge Discovery and Data Mining, 2004.

[11] Walaa Medhat, Ahmed Hassan, Hoda Korashy, "Sentiment analysis algorithms and applications: A survey", Ain Shams Engineering Journal, pp. 1093-1113, Volume 5, Issue 4, December 2014.

[12] Isa Maks, Piek Vossen, "A lexicon model for deep sentiment analysis and opinion mining applications",Decision Support Systems, pp. 680-688, 2012.