Sentiment Analysis and Machine Learning Approach: A Survey

Shubham Mishra [1], Vikas Pandey [2]
M. Tech Student [1] , Associate Professor (Dept. of Information Technology) [2]
Bhilai Institute of Technology, Durg, India [1,2]

Abstract: Sentiment analysis (SA) is an ongoing research area in the field of text mining. SA is the calculation of text perspective, emotions and subjectivity. This research paper provides a comprehensive overview of the latest updates in this area. In this study, we briefly surveyed and introduced a number of recently proposed algorithms and various SA applications. These articles are categorized according to their contribution to various SA technologies. Related areas of SA that have recently attracted researchers (transfer learning, emotional testing, and building resources). The main purpose of this survey is to provide near complete SA technology images and brief details on the relevant fields. The main contributions of this paper include the complex classification of a large number of recent articles, as well as an explanation of recent trends in emotion analysis and related fields.

Keywords: Sentiment analysis (SA) , Feature Extraction ,support vector machines (SVM), and naive Bayes, SentiWord Net

1. Introduction

Today, the Internet era has changed the way people to define their opinions. Currently, it is mainly done through blog posts, online forums, product review sites, and social media. Today, millions of people use social networking sites such as Facebook, Twitter, Google Plus, etc. to express their feelings, opinions, and share opinions about everyday life. Through the online community, we have an interactive medium where consumers can understand and influence others. Social media generates a lot of emotionally rich data in the form of tweets, status updates, blog posts, comments, comments and more. In addition, social media provides business opportunities by providing a platform to connect with customers for advertising. People rely heavily on user-generated online content to make decisions. For example, if someone wants to buy a product or use a service, you first review the comments online and discuss them in social media before making a decision. The amount of content generated by users is too large for general users to analyze. Therefore, automation is required, and various emotion analysis techniques are widely used.

Sentiment Analysis (SA) informs the user if the product information is satisfactory before purchase. Marketers and companies use this analytics data to understand their products or services and make them available on demand. Text information retrieval techniques are mainly focused on the processing, retrieval or analysis of the presence of fact data. Although facts have objective elements, there are other text representations that express subjective characteristics. These are mainly the ideas, emotions, ratings, attitudes, and emotions that form the core of emotion analysis (SA). This provides many challenging opportunities for developing new applications, mainly due to the proliferation of information available in online resources such as blogs and social networks. For example, recommendations for projects suggested by the recommendation system can be predicted by considering considerations, such as by using an SA for positive or negative opinions about these projects.

2. Background

The authors are investigating the direction of emotion analysis. This is a social factor to analyze scenes created after text statements generated through social media platforms. The important considerations of this white paper are described below.

[1]. Pandey *et al.* (2017) proposed to Sentiment analysis is an important of Data mining areas, including the identification and analysis of emotional content, are usually available on social media. Twitter is one of the social media that many users use on specific topics in tweets. These tweets can be analyzed in a cluster-based way to find user feedback and emotions. However, due to the nature of the Twitter data set, in this article, traditional clustering techniques have enhanced the method of emotion analysis, suggesting a new legacy (CSK) based on the discovery of K mean and cuckoo search. There are proposed methods used to find the best cluster head from the emotional content of the Twitter data set. Various Twitter data from the proposed procedure were tested on different twitter databases using different customizations, variable evolution, cuckoo search, modified cuckoo search, Gaussian cuckoo and two n- n-gram methods. Experimental results and data analysis proved that the proposed method is better than the current one. The proposed procedure implies a theoretical means for the future study of data created by social networks / media. This method is very effective in designing a system that can provide a closing review on any social issue.

[2]. Singh & Kumari (2016). Study of the prevalence of online social media and the use of short text messages are potential sources of sending sheep information, especially in emotions, so exceptional assessment and analysis is an important issue for current research. The main tasks in this area are the customization of noise, compatibility, emotions, economic agreement, and slang data. The task is to see the impact of typing on Twitter's statistics, especially in terms of slang terms. The proposed pre-processing method relies on the slang second-current word to check the meaning and emotional translation of the slang, and check its importance to find bound and conditional random fields. Gram is used. The slang experiment was conducted to follow the effectiveness of the experimental method that clearly reflects the accuracy of the emotion evaluation.

[3]. Kolchyna *et al. (*2015) this article introduces two methods of emotion analysis: i) Dictionary based method. ii) machine learning method. In this paper they describe some techniques for applying these methods and discuss how they could be used to rank sentiments in Twitter messages. They are doing a balanced study of the overall impact, the culprit, and the representation of various dictionary collections by social media, and the accuracy of the dictionary-based Twitter rating is the choice of practice generation and machine learning emotion rating Make it more important. You can emphasize the process. Run this algorithm on the SemEval--2014 Competitive Task 2-B quality Twitter database to reduce the performance of key emission analysis methods on Twitter. The results showed that the way to learn SVM and Navy Base Classifier machines is better than dictionary techniques. They provide a new combination of tools based on machine learning input methods as input methods. The combined procedure proved to be a more accurate assessment. They also show that cost-effective classic developers improve performance up to 7% up to the maximum database.

[4]. Kharde & Sonawane (2016) proposed to growth of Network technology development can be used in large data networks for Internet users and can generate large amounts of data. The Internet is a platform for learning online, changing ideas and providing feedback. Social networking sites such as Twitter, Facebook, and Google are becoming more popular because they can discuss topics, talk to different communities, and post messages around the world. Some surveys have been conducted in the field of sentiment analysis of Twitter data. This survey focuses on sentiment analysis of Twitter data. It can be used to analyze information in a tweet, whether it is very positive or negative, and in some cases moderate, where highly structured and diverse ideologies are either positive or negative, or in some cases it is neutral. This article provides existing monitoring surveys and comparative analysis, such as machine learning and dictionary-based methods, as well as matrices. Many machine learning algorithms, such as new boxes, max Entropy and support vector machines, have been used to find Twitter data streams. He also described the general challenges and emotions applied to Twitter.

[5].Cliché (2017) in this paper an attempt state-of-the-art Try to generate the most advanced Twitter emotional classifiers using convolutional neural networks (CNN) and long-term short-term memory (LSTM) networks. Our system pretrains word embedding using a large amount of unlabeled data. Then they use subsets to use unlabeled monitoring to pre-train the embedding. The last CNN and LSTM were trained on the SemEval-2017 Twitter data set, where the embedding was fined again. They have combined several CNNs and LSTMs to improve performance. Our approach is ranked first among the 5 English subtasks of the 40 team.

[6]. Pang & Lee (2002) proposed to Analyze Twitter's **sentiment** to monitor real-time awareness of related products and events related to events. The first step in emotion analysis is text preprocessing of Twitter data. Most of Twitter's research on emotion analysis focuses on the extraction of new emotional features. However, the selected preprocessing method is ignored. This paper describes the effects of text pre-processing on the performance of emotional classification in two types of classification tasks, and the classification of six pre-processing methods using two feature models and four classifiers on five Twitter data sets Discuss They summarized the performance. Experiments show that preprocessing methods to replace negative words with expanded abbreviations improve the accuracy and F1 metrics of the Twitter emotional classification classifier, but delete URLs, delete numbers, or stop words showed that. Naive Bayesian and random forest classifiers are more sensitive than logistic regression, and support vector machine classifiers when different preprocessing methods are applied.

[7]. Amolik *et al.* (2016) study to Focusing on emotional analysis and text perspectives. Emotion analysis can be called opinion mining. Emotion analysis finds and demonstrates how people feel about a particular source of content. Social media includes a variety of emotional data such as tweets, blogs, post updates, posts, and more. Emotion analysis of this large amount of generated data is very useful for expressing public opinion. Compared to extensive emotion analysis, Twitter's emotion analysis needs attention because it has slang, spelling errors, and repetitive roles. The maximum length of each Tweet on Twitter is 140 characters. Therefore, it is important to determine the correct mood of each word. Our project presents a very accurate emotion analysis model and the latest comments on upcoming Bollywood and Hollywood movies. By classification of eigenvector machines, support vector machines, Naïve Bayes, etc., these positive, negative and neutral are classified correctly to express the emotion of each tweet.

[8]. Giachanou & Crestani (2016) proposed to Emotion analysis on Twitter is an important area of recent research. Twitter is one of the most popular microblog platforms, allowing users to post their thoughts and opinions. Twitter's emotional analysis solves the problem of analyzing tweets based on their opinion. This survey investigates and briefly describes the emotional analysis algorithm on Twitter and gives an overview of the topic. The studies presented are classified according to the method they follow. It also covers the areas of Twitter's analysis of emotions, such as Twitter opinion search, tracking emotions over time, satire detection, emotion detection, and tweet emotion quantification. The main contribution of this survey is a demo that recommends the Twitter sentiment analysis method. It is classified by the technology used. The latest research trends in topics and related fields were also discussed.

[9]. Saif *et el.* (2016) study on Most of the work on tweet sentiment analysis is single language, and the models generated by machine learning strategies are not extensions between languages. Cross-language sentiment analysis is usually performed by machine translation methods that translate a given source language into a selected target language. Machine translation is expensive and the results provided by these strategies are limited by the quality of the translation. This article introduces language-independent translation methods for Twitter sentiment analysis, and points out the correct poles of various (or multiple) languages. The proposed method requires less parameters to learn, while tweets using deep convolution neural networks with character level embedding are more accurate. The resulting model provides an easy way to learn the basics of all the languages used in the training process without completing the translation process. Based on tweets from four different languages, they evaluated the effectiveness and effectiveness of the proposed method based on our experience with the tweet corpus and showed that the method outperformed the baseline. In addition, they visualize the knowledge gained in our way and qualitatively validate the emotional classification of their tweets.

[10]. R. Quirk *et el.* (1985). Study on to Emotion analysis on Twitter has recently gained much attention due to its widespread use in the commercial and public sectors. In this article, we introduce the dictionary-based sentiment analysis method SentiCircles on Twitter. Senti Circles takes into consideration the various contextual word patterns of the current tweet, as opposed to the typical dictionary-based method that provides a fixed and static prior emotional polarity, regardless of background. Update pre-specified strength and polarity in combination with. And the corresponding one. Our approach allows for detection of emotion at entity and tweet levels. They used three different emotion data to evaluate the methods presented in the three Twitter data sets and derive deductive emotions for the word. The results show that our method is significantly superior to baseline accuracy and subjective (native and polar) and polar (positive and negative) detection at the physical level. In both datasets, tweet level sentiment detection performance outperforms the highest Senti-Strength accuracy of 4% to 5%, but in the third data set, F-measure slightly exceeds 1%.

[11]. Ghiassi *et el.* (2016) proposed on to Social media provides valuable feedback about a company's brand. They use Active Skills Monitoring Engineering and Synthetic neural Network Dynamic Architecture to provide Twitter's sentiment analysis brand oriented approach. The proposed method solves the distribution issues associated with typical features and brands associated with the Twitter language. Effective performance of Twitter datasets with two unique brands. The technical features of the brand are the final features of the 7-dimensional amplitude with only the final feature density. Reduce the complexity of the classification problem to reduce the dimensionality of the representation. Two sets were used for the third and fifth tweets emotion types of each brand. They reviewed five categories and expressed moderate feelings of particular interest to companies and brand management staff. They compare the proposed approach with the performance of two advanced Twitter passion analysis systems from the educational and commercial fields. The results show that it is superior to the state-of-the-art system, the F1 measurement rating is up to 88%, and expresses a clear and consistent feeling. Furthermore, they are a feature of Twitter and are very effective at controlling the metabolism of Twitter's emotional expressions, but there are only seven dimensions. The monitoring feature applies to most brand recommendations, and most features within most features allow the staff of the researcher or brand manager to quickly direct other brands to twitter to create highly efficient tights. Feature representation for passion analysis.

[12]. Jianqiang *et al.* (2018) propose Through Twitter sentiment analysis, organizations can investigate the general sentiment of related events and products. Most of the research has focused on obtaining emotional features by analyzing lexical and syntactic features expressed in emotional words, pictograms, exclamation marks and the like. This article describes the embedding of words obtained by unsupervised learning with a potentially large corpus. Contextual semantic relationships and co-occurrence statistical features between words embedded in these words are combined with n-gram features and word affective polarity score features to form a set of tactile emotional features. This feature set is integrated into a deep convolution neural network for training and predicting emotion classification tags. The word n-gram model of the five Twitter data sets is used as a baseline model to show Twitter's emotional classification exactitude and F1-Measure of Twitter- sentiment classification.

3. SENTIMENT ANALYSIS

Sentiment analysis could be defined as a process of automatically mining attitudes, opinions and emotions from text, speech, tweets, and database sources through natural language processing (NLP). Emotion analysis classifies ideas in the text as "positive" or "negative" or "neutral". Also known as subjective analysis, opinion mining, and evaluation extraction. The terms opinion, emotion, opinion, and belief are used interchangeably, but with some difference between shows below.

- ***Opinion**:* A conclusion open to dispute (because different experts have different opinions)
- ***View:*** subjective opinion
- ***Belief:*** deliberate acceptance and intellectual assent
- ***Sentiment:*** opinion representing one's feelings

3.1 Pre-processing of the datasets

Tweets have many opinions about the data. They are expressed differently for each user. The tweet datasets used in this study are marked as two categories. Because of negative polarity and positive polarity, sentiment analysis of the data makes it easier to observe the effects of different functions. Raw data with polarity is very susceptible to inconsistencies and redundancy. The pre-processing of tweets has the following points.

- Remove all URLs (e.g. www.xyz.com), hash tags (e.g. #topic), targets (@username)
- Correct spelling ,Process repeated character sequences
- Replace emoticons with your own emotions.
- Remove using all the symbols, punctuations and numbers.
- Remove Stop Words
- To define the Acronyms (we can use a acronym dictionary)
- Delete non-English tweets

3.2 Feature Extraction

Preprocessed data sets have many unique characteristics. Feature extraction extracts aspects from the processed data set. Later, this aspect was used to calculate the positive and negative polarity in the sentence. This is useful for determining individual opinions using models such as unigram and bigram [13]. Machine learning techniques need to represent key features of text or documents for processing. These important features are considered feature vectors for classification work. Some sample structures that have been reported in the literature are:

1. Words and Their Frequencies:

Unigram, bigram, and n-gram models, including frequency numbers, are considered features. Research on the use of the presence of words rather than frequencies is increasing to better explain this function. Pang et al. [14] shows better results by using presence instead of frequency.

2. Parts of Speech Tags

Adjectives, adverbs, verbs and nouns are good indicators of subjectivity and emotion all these are part of speech. They could generate syntax dependency patterns through parsing or dependence of trees.

3. Opinion Words and Phrases

In addition to specific words, they can use several phrases or idioms to convey emotions as features. For example, let's make someone an arm and a leg.

4. Position of Terms

The position of a term in the text can affect the extent to which the term differs from the overall atmosphere of the text.

5. Negation

Negation is an important but it is difficult to explain it. The existence of negation usually changes the polarity of the view.

6. Syntax

Syntactic patterns such as collocations are to use by many researchers as a feature to learn subjective patterns.

3.3 Training

Supervised learning is an important technique for solving of sorting the problems. Training a classifier makes it easier to predict future unknown data.

3.4 Classification

3.4.1 Naive Bayes:

It is a probability classifier that can learn to examine patterns of groups of classified documents [9]. Compare the content to a list of words to categorize the document into the correct categories or categories. Given a tweet, c * is a class assigned to d, a probability classifier that can learn to examine patterns of groups of classified documents [9]. Compare the content to a list of words to categorize the document into the correct categories or categories. Let d be a tweet and c * be a class assigned to "d" here,

$$C^* = \arg mac_{c^{P_{NB}}} (C|d)$$

$$P_{NB}(C|d) = \frac{(P(C)) \sum_{i=1}^{m} p(f|c)^{n_{i(d)}}}{P(d)}$$

According to the above equation, "f" is a "feature" and the number of features (fi) is represented by ni (d) and is present in d representing a tweet. Here, m means no. function. The parameters P (c) and P (f | c) are calculated by maximum likelihood estimation and smoothing is used for the invisible features. The Python NLTK library can be used to train and classify using Naïve Bayes machine learning technology.

3.4.2 Maximum Entropy

The maximum entropy classifier makes no assumptions about the relationship between features extracted from the data set. The classifier always tries to maximize the entropy of the system by estimating the conditional distribution of class labels.

Maximum entropy deals with even overlapping features and is the same as logistic regression to find a distribution on a class. Conditional distributions are defined so that MaxEnt does not make independent assumptions about its properties like Naive Bayes.

The model is represented as follows:

$$P_{ME}(c|d,\lambda) = \frac{exp[\sum_i \lambda_i f_i(c,d)]}{\sum_c exp\,[\sum_i \lambda_i f_i(c,d)]}$$

3.4.3 Support Vector Machine:

Support vector machines analyze data, define decision boundaries, and use the kernel for calculations performed in the input space [15]. The input data is two sets of vectors of size m. And each data expressed as a vector is classified into one class. Second, I found that the gap between the two classes goes far beyond the other documents. Distance defines the margin of the classifier and maximizing the margin reduces the hesitation decision. SVM also supports classification and regression to help statistical learning theory. It also helps to accurately identify the factors that need to be considered in order to understand it correctly.

4. Machine Learning Approaches

Machine learning based on the methods to use classification the techniques to classify text into classes. There are two main types of machine learning technology.

4.1 Unsupervised learning:

It relies on clustering because it contains no categories and no correct target is provided.

4.2 Supervised learning:

Because this is based on a tagged data set, tags are provided to the model in the process. These tagged data sets are trained to experience meaningful output during the decision making process. The success of these two learning methods mainly depends on the selection and taking out of specific feature sets to detect emotions.

Machine learning methods for emotion analysis are mainly classified as supervised. Machine learning technology requires two sets of data.
1. Training Set
2. Test Set.
Many machine learning techniques have been developed to classify tweets as classes. Machine learning technologies such as Naive Bayes (NB), Maximum Entropy (ME) and Support Vector Machine (SVM) have been extremely successful in emotion analysis. Machine learning starts with the collection of training data sets. Next, they train the classifier on the training data. Once a supervisory classification technique is selected, important decisions are made to select features. They can tell us how the file is represented.

The most common features of sentiment classification are:

- Term presence and their frequency
- Part of speech information
- Negations
- Opinion words and phrases

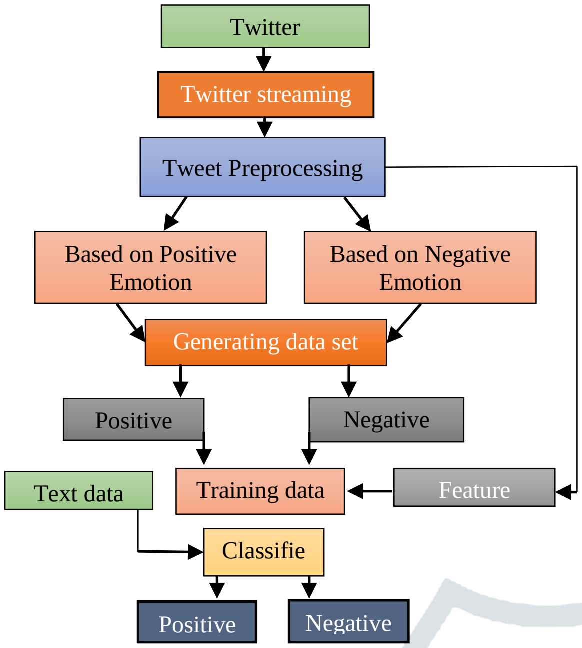


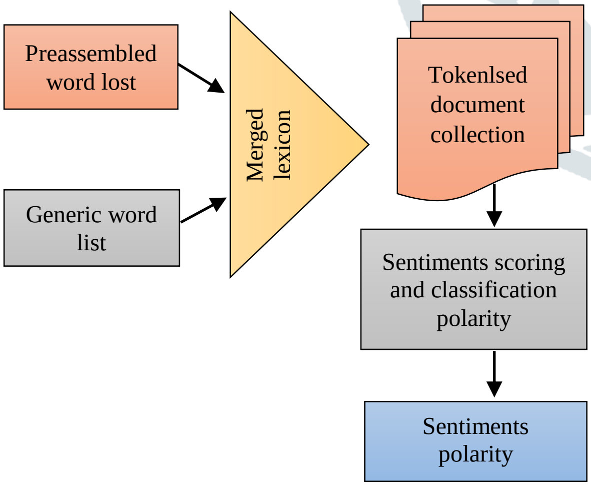
Fig.2

Sentiment Classification Based on Emoticons

For surveillance technology, support vector machines (SVM), and naive Bayes, maximum entropy is part of the most commonly used technology. If it is not possible to have an initial set of label documents / opinions to classify the remaining items, semi-supervised and unsupervised techniques are suggested.

4.3 Lexicon-Based Approaches

The dictionary-based approach [16] uses an emotional dictionary containing opinion words and matches it to data to determine polarity. They assign opinion scores to positive, negative, and objective opinions that explain the words contained in the dictionary. The dictionary-based approach relies mainly on pre-edited emotional words, phrases, and also an emotional glossary, also known as an idiomatic phrase developed for traditional propagation types such as the opinion Finder dictionary.


Preassembled word lost
Generic word list
Merged lexicon
Tokenlsed document collection
Sentiments scoring and classification polarity
Sentiments polarity

4.3.1 Dictionary-based:

This is usually based on the use of manually collected and annotated terms (seeds). Collections are added by searching dictionary synonyms and antonyms. An example of this dictionary is Word Net. This is used to develop a thesaurus called SentiWord Net.

Drawback: Can't deal with domain and context specific orientations.

Drawback: It cannot handle domain and context specific instructions.

4.3.2 Corpus-Based:

The purpose of the corpus-based approach is to provide a dictionary that is relevant to a particular area. These dictionaries are generated by a series of seed opinion terms that grow by searching for related terms using statistical or semantic methods. Statistics-based approach: latent semantic study (LSA). Semantic-based methods (using synonyms and antonyms, relationships from a thesaurus like WordNet, etc.) it can present also be an interesting solution.

5. Conclusion

In this article, they have first to propose a detailed procedure for performing the emotion analysis process to categorize Twitter's unstructured data into positive or negative categories. Next, they explained various techniques, like a knowledge-based technology and machine learning techniques, for sentimental analysis of Twitter data. In addition, they present a comparison of the parameters of supervised machine learning techniques discussed based on our determined parameters. Various techniques applied to emotion analysis are known to be domain specific and language specific.

References:

1. Pandey, A. C., Rajpoot, D. S., & Saraswat, M. (2017). Twitter sentiment analysis using hybrid cuckoo search method. *Information Processing & Management*, *53*(4), 764-779.

2. Singh, T., & Kumari, M. (2016). Role of text pre-processing in twitter sentiment analysis. *Procedia Computer Science*, *89*, 549-554.

3. Kolchyna, O., Souza, T. T., Treleaven, P., & Aste, T. (2015). Twitter sentiment analysis: Lexicon method, machine learning method and their combination. *arXiv preprint arXiv:1507.00955*.

4. Kharde, V., & Sonawane, P. (2016). Sentiment analysis of twitter data: a survey of techniques. *arXiv preprint arXiv:1601.06971*.

5. Cliche, M. (2017). BB_twtr at SemEval-2017 task 4: twitter sentiment analysis with CNNs and LSTMs. *arXiv preprint arXiv:1704.06125*.

6. Jianqiang, Z., & Xiaolin, G. (2017). Comparison research on text pre-processing methods on twitter sentiment analysis. *IEEE Access*, *5*, 2870-2879.

7. Amolik, A., Jivane, N., Bhandari, M., & Venkatesan, M. (2016). Twitter sentiment analysis of movie reviews using machine learning techniques. *International Journal of Engineering and Technology*, *7*(6), 1-7.

8. Giachanou, A., & Crestani, F. (2016). Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys (CSUR)*, *49*(2), 28.

9. Wehrmann, J., Becker, W., Cagnini, H. E., & Barros, R. C. (2017, May). A character-based convolutional neural network for language-agnostic Twitter sentiment analysis. In *2017 International Joint Conference on Neural Networks (IJCNN)*(pp. 2384-2391). IEEE.

10. Saif, H., He, Y., Fernandez, M., & Alani, H. (2016). Contextual semantics for sentiment analysis of Twitter. *Information Processing & Management*, *52*(1), 5-19.

11. Ghiassi, M., Zimbra, D., & Lee, S. (2016). Targeted twitter sentiment analysis for brands using supervised feature engineering and the dynamic architecture for artificial neural networks. *Journal of Management Information Systems*, *33*(4), 1034-1058.

12. Jianqiang, Z., Xiaolin, G., & Xuejun, Z. (2018). Deep convolution neural networks for Twitter sentiment analysis. *IEEE Access*, *6*, 23253-23260.

13. Socher, Richard, et al. "Recursive deep models for semanticcompositionality over a sentiment Treebank." Proceedings of theConference on Empirical Methods in Natural Language Processing (EMNLP). 2013.

14. Pang, B.and Lee, L. “A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts”. 42nd Meeting of the Association for Computational Linguistics[C] (ACL-04). 2004, 271-278.

15. Liu, S., Li, F., Li, F., Cheng, X., &Shen, H.. Adaptive co-training SVM for sentiment classification on tweets. In Proceedings of the 22nd ACMinternational conference on Conference on information & knowledgemanagement (pp. 2079-2088). ACM,2013.

16. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., &Stede, M..“Lexicon basedmethods for sentiment analysis”. Computational linguistics, 2011:37(2), 267-307.