Review of Computer Engineering Research
2019 Vol.6, No.2, pp. 84-91
ISSN(e): 2410-9142
ISSN(p): 2412-4281
DOI: 10.18488/journal.76.2019.62.84.91
© 2019 Conscientia Beam. All Rights Reserved.

A SURVEY ON SENTIMENT ANALYSIS ALGORITHMS AND DATASETS

Reena G.Bhati

Computer Science Department, Tilak Maharashtra Vidyapeeth, Pune, India.
Email: reena4bhati@gmail.com Tel: 9850285095

ABSTRACT

In this paper we present a deep literature review on existing system for sentimental analysis. Basically sentimental analysis (SA) is the measurement of preference of people's thoughts via natural language processing. The main aim of sentiment analysis is to know the orientation of the sentiment described in script. In recent decades the researcher focuses on the study various algorithms for relevant research results of the sentiment analysis. This research paper provides a comprehensive overview of this field's latest update. In this review, some recent proposed improvements of algorithms and various SA applications are explored and briefly described. The aim of this paper is to provide knowledge about the different method related to sentimental analysis also how they are classified, what the applications of this analysis.

Article History
Received: 6 August 2019
Revised: 10 September 2019
Accepted: 14 October 2019
Published: 2 December 2019

Keywords
Data mining
Natural language processing (NLP)
Sentimental analysis (SA)
Convolutional neural network (CNN) unsupervised learning
Text mining
Feature representation.

Contribution/Originality: This study contributes to the existing literature by studying the existing systems and showing the disadvantages and comparative analysis based on various parameters.

1. INTRODUCTION

The information sharing between a sender & receiver for exchanging data are becoming popular trend in young generation [1]. Due to easily availability of web connectivity, sharing personal opinions with others are increased. Opinion is nothing but the sensation or feeling or behaviour of the person towards certain topics, as these centuries show a great deal of concern in sharing the material on the internet. Because of this it is becoming common nowadays to communicate & swap their views & views with others. Also exchanging information are used for defining the rate of product, creating different kinds of blogs on particular article or issues. By taking example for buying any product from online shopping, customer reviews the rating for that particular product & then decides to buy the product. The opinion of person represents the view of that product. Similarly, emotion of person correlates with is behaviour related to the situation & that emotion known as sentiment. Sentiment analysis is therefore basically a class of natural language processing to trace the community's behavior toward a specific item or topic [2]. Generally the Sentiment classification is having two classes named as positive & negative. The predefined data set positive & negative reviews can be identified by taking review of ratings of online which are pre defined as 1–5 stars. For example the star rating with 4 & 5 star reviews are basically treated as positive review, & 1 & 2 star reviews are treated as a negative review [3]. With the help of recent studies of sentiment analysis, the methods of SA are divided into two viewpoints. The first are based on the granularity of text analysis, which are again classified into three levels i.e. aspect level, document level & statement level. The remaining one are based on the method of operation which are having three types: deep learning, rule-based, & machine learning [4]. The Aspect-level SA objective of classifying feeling with regard to the particular elements of organizations. Document-level SA is a feeling or a positive or negative opinion used to identify the expression of an opinion document. Initial stage is to identify when the phrase is subjective or objective to define the phrase rank SA [5, 6]. The goal of Sentence-level SA is to identify the expression behind in each sentence [7]. There are no. of challenges in sentiment analysis. An opinion word can be assumed as positive in one situation or as negative in another situation. At that time the way of opinion is totally dependent on situation. As the way of thinking of people is totally defers as it changes from person to person. Some people carry their opinions immediately while other people don't express opinions immediately. Hence to analyzing sentences one at a time may become difficult because of reviews of peoples opinion will have both positive & negative attitudes.

The rest of the paper is organized into four sections. The section I already given the introduction about Sentimental Analysis. Section II gives the details about types of sentimental analysis. Section III describes the overview of literature survey. Section III demonstrates the proposed methodology. Section IV gives the idea about different dataset used in SA. Section V gives the conclusion of proposed methodology.

2. TYPES OF SENTIMENT ANALYSIS

There are basically three primary methods for analyzing sentiments: language assessment centered on machine learning and lexicon based [8, 9]. But to categorize the SA assessment, i.e., mostly two methods are used. Approach based on machine learning and lexicon. To determine the polarity, machine learning-based approach is introduced to classify the writing & polarity Lexicon-based technique. Sentiment dictionary with view phrases and combine them with information are used in Lexicon-based technique. Here the feeling results are assigned as how the terms found in the dictionary are positive, negative and objective to the statements of opinion [10].

The linguistic approach is used with lexicon based methods. This approach is used to define the syntactic characteristics of the phrases or words, the negation & the structure of the text [2]. Figure 1 depicts the types of sentiment analysis in tree structure. Which are categories in machine language, linguistic analysis and lexicon based approach.

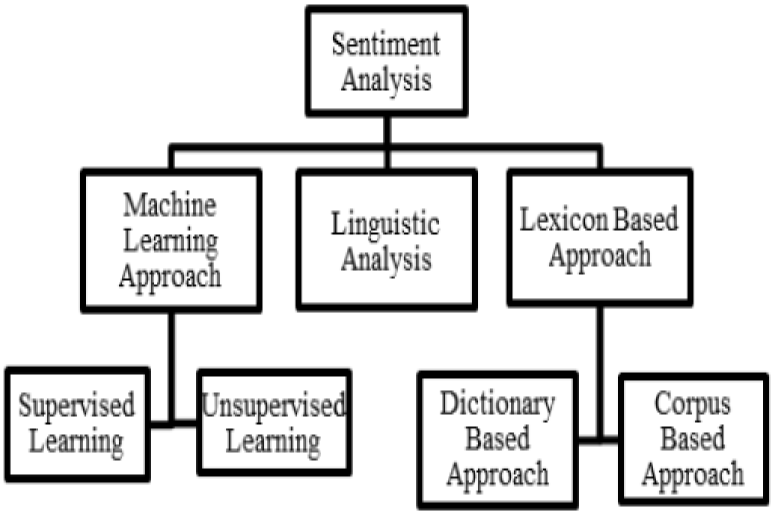

Figure-1. Types of SA [2].

Source: Lincy and Naveen [2].

Machine training techniques can again be classified into monitored and unsupervised techniques of teaching. Text classification classifies keyword-based word, these keywords are basically subject-related, e.g., athletics, polity, and products. Words of view are essential in the ranking of feelings, indicating favorable or negative views, e.g. excellent, better, poor, bad, ok, bad, best, etc. All the current monitored techniques of teaching have been introduced to issues with text classification. Various classification models such as Support Vector Machine (SVM), classification of Naïve Bayes (NB) have been used [3]. In monitored techniques, a enormous amount of labeled teaching papers are suggested. Using training documents, unsupervised methods are used when it is difficult to find these labeled unsupervised methods. Words of feeling are essential for ranking of feelings in a document or message. Identifying terms of feeling is simple, terms of feeling are very helpful for classifying feelings in unsupervised learning [8, 9]. Here the ranking requires position to recognize some set syntactic models; these models are lastly used to convey specific background view. Syntactic patterns are composed with POS tags [3]. As Approach based on Lexicon is used to examine writing according to its view. The Lexicon-based approach is based on a lexicon of feelings, a set of concepts of recognized & precompiled feelings. This approach is divided into a dictionary-based strategy based on technique & corpus. Both techniques use statistical or semantic techniques to discover polarity of feelings. Dictionary-based strategy is used to check for synonyms & antonyms to find emotional phrases ' view. While the corpus-based strategy uses the roster of phrases to find individuals ' view linked to a particular situation. Statistical or semantic techniques are used for this method [7].

3. RELATED WORK

A subdivision of text mining called as Sentiment analysis, in which computer technology are used for extracting & categorize emotions in text for this purpose natural language processing are used [4]. Sentiment analysis has a very wide range of applications. Nowadays, everyone is addicted to social media for showing what's going their life. As social media is plays important role in showing peoples opinion towards any particular event or product. Hence social media is favorite research topic for researcher & advantage of this platform is no. of user are unlimited. So it is best application of sentimental analysis. Figure 2 shows some research works regarding sentimental analysis done by different researchers.

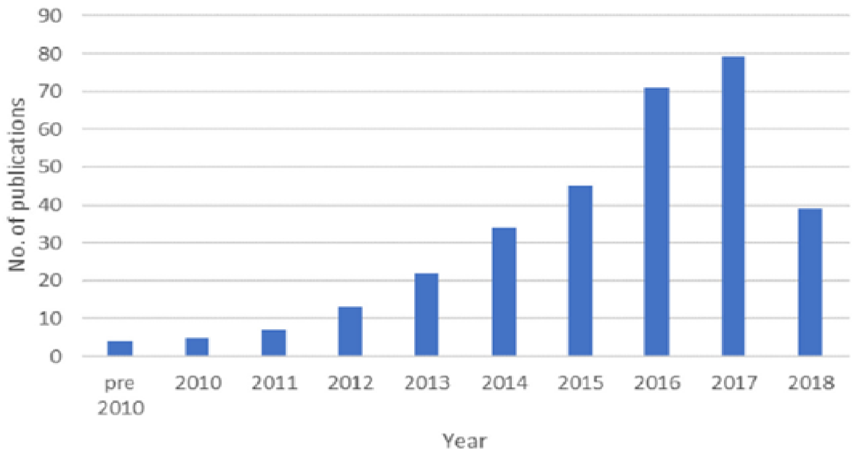

Figure-2. Graphical view of research paper year wised published by various researcher [11].
Source: Al-Ayyoub, et al. [11].

The above diagram shows the research paper are published on sentimental analysis as per survey of Al-Ayyoub, et al. [11]; Archana and Deipali [12]; Erion and Maurizio [13]; Diksha, et al. [14].

This study provides a thorough summary of the work conducted on Arabic SA. The study also describes the organizations released documents centered on issues linked to the SA. Trying to define the holes for potential studies in this sector in the current literature setting footprints. Kharde and Sonawane [15] conducts studies on the methods of emotional assessment by using Twitter information as an instance of feedback. In their research study lexicon method & machine method are compared with existing methods & gives overview about different applications & challenges faced by different authors.

As per research by Yang, et al. [16] the lexicon-based type SA required compulsorily sentiment dictionary but it is simple to handle. The researcher uses an improved SOPMI algorithm for computing & emotional vocabulary modeling which is more effective method of this type of SA. In the review of article done by Weiss, et al. [17] gives a detailed survey of transfer learning & the related methods with applications. According to survey based on different researchers, the Weiss et al. found different method of transfer learning.

These method classified into homogeneous & heterogeneous methods & collect information about the negative transfer problem. Semantic polarity algorithm are used by researcher Turney and Littman [18]. This algorithm are mostly in the lexicon based approach to analyzing the text sentiment tendency. The result will be in textual data format with validated effective & appropriate datasets.

Routray, et al. [19] explains machine learning techniques in the efficient way to analyze the sentiment laden terms in a document. the researcher discussed the challenges of SA like subjectivity classification, word sentiment classification, document sentiment classification & opinion extraction. According to the learning method of researcher Pan and Yang [20] the transfer learning is categorized into four approaches, named as relational knowledge transfer approach, feature based representation transfer approach, instance transfer approach & parameter-transfer approach. In Instance-transfer method, the exchange of information are done through the source & target domain. The data are sorted out can from source domain by re-weight. Because of this unwanted data are eliminated from target domain [21]; [22]. Where in feature based representation transfer approach, the target & source domain required partially limited features. It is mandatory to have same features space between them while data are transferred & then performed machine learning. Parameter-transfer method controls the parameter sharing between the source & target domains. With the help of parameter transfer method, large number of datasets are transfer from trained model to the target task.

Yang, et al. [23] invented a method for document classification using two level hierarchical attentional network. The levels are used for this method is word & sentence. This method gives the pictorial view of specific words selected by the attention layer. This will increase capability of attention mechanism towards the importance of classification category of individual words & sentences.

Vieriu, et al. [24] gives the information about a boosting oriented transfer learning method. This method share the data to various target areas by reducing the cost of labeled target areas by field by learning training samples from multiple fields. In past few decades the Convolutional Neural Network (CNN) are used in NLP for text analysis which gives the improved & effective results in computer vision applications. By taking reference of this many researcher do their research study on this method. Kim [25] proposes the simple CNN model with a small amount of hyper parameters combined with static word vectors. These word vectors are used to classify at sentence level for obtaining good results.

McCann, et al. [26] and Peters, et al. [27] proposes a method of merging embedding from other tasks with different levels of input data. The neural network is used for obtain output via English-German translation task called as context vectors (CoVe).These are used for defining classification of problem, understanding of text, semantic sentiment analysis via NLP.

The new architecture developed by Conneau, et al. [28] named as VDCNN which is used for text analysis. Basically in this architecture the deeper CNN having 29 convolutional layers are applied to NLP. The model deep residual network (ResNet) combines with VGGNet (Visual Geometry Group). Johnson and Zhang [29] uses the Long Short-Term Memory (LSTM) which more difficult embedding method. This method uses the variable sized text regions for obtaining the results by combing convolutional layers & LSTM-style region embedding. The effect of LSTM combined with CNN are more the advantageous than the regional embedding method. Following table give the short explanation about above literature review.

Dai and Le [30] narrates the methods for sequence learning with recurrent network. For this fine-tuning of the language model (LM) method are used with the help of unlabelled data. The purpose of this method is to forecast the subsequent content in the sequence & to select the input sequence into the vector & predict the input again by using sequence-auto encoder. This approach improves the strength of training about LSTM & become advantageous for unsupervised & supervised learning. By taking reference study of Howard and Ruder [31] the Universal Language Model Fine-tuning (ULMFiT) are used for NLP. This method recommends training language models for semantic information & becomes the backbone of the classifier by giving great results for various NLP tasks. Table 1 shows the list of different approaches and their accuracy with the list of dataset.

Table-1. Different approaches by different researchers [2].

Author	Basic concept	Dataset	Accuracy
Conneau, et al. [28]	VDCNN	Yelp	95.72
Kim [25]	CNN	MPQA/SST-2	89.60/88.10
Johnson and Zhang [29]	LSTM	IMDB	94.06
McCann, et al. [26]	LSTM	IMDB/SST-5	91.80/53.70
Peters, et al. [27]	Bidirectional LSTM	SST-5	54.70
Dai and Le [30]	LSTM	IMDB	92.80
Howard and Ruder [31]	AWD-LSTM	IMDB	95.40

Source: Liu, et al. [4].

4. DATASETS FOR SENTIMENT ANALYSIS

The model or algorithm's legitimacy & precision can only be efficiently validated by choosing the suitable datasets. How to pick the finest data set for experiment in the longest moment is a valuable issue. Hence, evaluating user-friendly databases selects more appropriate dataset for validation of technique. Following is the dataset used for their study by Liu, et al. [4].

i. IMDB

In the 2011 ACL document, which is a information collection for binary emotion classification, the big movie assessment dataset IMDB [32] was suggested, each sample being a txt, folder, including instruction cards, test panels, and no marked information points. There are 25000 pairs of practice, 12500 favorable & adverse & 25000 pairs of testing, 12500 favorable & adverse.

ii. Stanford Sentiment Treebank

Stanford Sentiment Treebank [33] is Stanford University's linguistic lexical dataset, including a fine-grained mental tag of 215154 phrases in a 11855-phrase parse tree. It is split into two functions, one is a two-category job, including 6920 instruction kits, 872 validation sequences, and 1821 sample sequences; one is a five-category job, containing 11855 sentences and 215154 sentences (c).

iii. Yelp

The Yelp dataset is an inner dataset released by Yelp, America's biggest evaluation platform. This dataset is a subgroup of Yelp & coated dealers, recommendations, & user information for private, instructional, & academic reasons. Yelp expects that this dataset will be used by more academics to create study more interesting. The dataset can be used in three ways below. The first is the ranking of images. The second is the handling of natural language. In the customer assessment dataset, there are a number of mining metadata that can be used to infer semance, company characteristics, and feelings. The objective is the mining of images. Mining the connection between customers, for instance, to discover laws of use.

iv. Multi-Domain Sentiment Dataset

In the 2007 ACL article, which includes brand feedback from a number of distinct fields obtained from Amazon, John Blitzer suggested Multi-Domain Sentiment Dataset [34]. Com. Comments have concentrations (1 to 5 concentrations) and can be transformed to binary labels. The information collection includes more than 100,000 phrases that can be split into favorable & adverse classifications or five classifications of powerful favorable, soft positive, neutral, weak adverse, powerful negative.

v. Sentiment140 (STS)

Sentiment140 [35] obtained 1600,000 tweets from the api twitter. These tweets were annotated (Zero, 2 & 4 are negative, neutral & good, respectively). The tweets contain the following six fields: target, tweet polarity; ID, tweet ID; location, tweet location; flag, request; tweet device; sms, tweet material. Table 2 depicts the various dataset and their comparative study with classification types and short description.

Table-2. Comparative study of dataset [4].

Dataset	Classification	Number	Description
IMBD	Two category	100000 sentences	More review
Stanford sentiment treebank	Two category & five category	11855 Sentences divided into 215154 phrases	-
Yelp	-	More than 163 million pieces of data	User review, business information
Multi domain sentiment dataset	Two category & five category	More than100000 sentences	Multi domain product review
Sentiment 140	Three category	1600000 Tweets	Twitter information

Source: Liu, et al. [4]

The above table gives the comparative study about all five dataset listed by author [4].

5. CONCLUSION

This review gives a discussion & complete analytical study on sentiment analysis in different directions. In this study, different method regarding transfer learning & their different approaches effect of negative transfer problem are carried out. In the decision making process while selection about product, service, movie, social issues & observation carried out by various researcher sentimental analysis plays important role.

Funding: This study received no specific financial support.
Competing Interests: The author declares that there are no conflicts of interests regarding the publication of this paper.

REFERENCES

[1] M.-H. Chen, W.-F. Chen, and L.-W. Ku, "Application of sentiment analysis to language learning," *IEEE Access*, vol. 6, pp. 24433-24442, 2018. Available at: https://doi.org/10.1109/access.2018.2832137.

[2] W. Lincy and K. M. Naveen, "A survey on challenges in sentiment analysis," *International Journal of Emerging Technology in Computer Science & Electronics*, vol. 21, pp. 409-412, 2016.

[3] P. Chiranjeevi, D. T. Santosh, and B. Vishnuvardhan, *Survey on sentiment analysis methods for reputation evaluation. In cognitive informatics and aoft computing*. Singapore: Springer, 2019.

[4] R. Liu, Y. Shi, C. Ji, and M. Jia, "A survey of sentiment analysis based on transfer learning," *IEEE Access*, vol. 7, pp. 85401-85412, 2019. Available at: 10.1109/ACCESS.2019.2925059.

[5] P. Priyanka and Y. Pratibha, "Sentiment analysis levels and techniques: A survey," *International Journal of Innovations in Engineering and Technology*, vol. 6, pp. 523-528, 2016.

[6] D. Abdullah and J. Anurag, "Survey paper on sentiment analysis: In general terms," *International Journal of Emerging Research in Management &Technology*, vol. 5, pp. 1093-1113, 2014.

[7] M. Walaa, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, pp. 1093-1113, 2014. Available at: https://doi.org/10.1016/j.asej.2014.04.011.

[8] D. M. E.-D. M. Hussein, "A survey on sentiment analysis challenges," *Journal of King Saud University-Engineering Sciences*, vol. 30, pp. 330-338, 2018.

[9] M. Sadegh, R. Ibrahim, and Z. A. Othman, "Opinion mining and sentiment analysis: A survey," *International Journal of Computers & Technology*, vol. 2, pp. 171-178, 2012.

[10] A. M. Dudhat, R. R. Badre, and K. Mayura, "A survey on sentiment analysis and opinion mining," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 2, pp. 6633-6639, 2014.

[11] M. Al-Ayyoub, A. A. Khamaiseh, Y. Jararweh, and M. N. Al-Kabi, "A comprehensive survey of arabic sentiment analysis," *Information Processing & Management*, vol. 56, pp. 320-342, 2019. Available at: https://doi.org/10.1016/j.ipm.2018.07.006.

[12] S. Archana and G. Deipali, "Sentiment analysis and challenges involved: A survey," *International Journal of Science and Research*, vol. 4, pp. 1928-1932, 2015.

[13] C. Erion and M. Maurizio, "Word embeddings for sentiment analysis: A comprehensive empirical survey," *arXiv:1902.00753v1*, *DBLP:journals/corr/abs-1902-00753*, *CoRR* , *abs/1902.00753*, 2019.

[14] S. Diksha, G. Shubham, J. Joy, and M. Richa, "Sentiment analysis," *International Journal of Engineering, Science and Mathematics*, vol. 8, pp. 46-52, 2019.

[15] V. A. Kharde and P. S. Sonawane, "Sentiment analysis of twitter data: A survey of techniques. arXiv:1601.06971. Available: https://arxiv.org/abs/1601.06971 " 2016.

[16] A. M. Yang, J. H. Lin, Y. M. Zhou, and J. Chen, "Research on building a Chinese sentiment lexicon based on SO-PMI," *Applied Mechanics and Materials*, vol. 263, pp. 1688-1693, 2013.

[17] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 1, pp. 1-40, 2016.

[18] P. D. Turney and M. L. Littman, "Measuring praise and criticism: Inference of semantic orientation from association," *ACM Transactions on Information Systems*, vol. 21, pp. 315-346, 2003. Available at: https://doi.org/10.1145/944012.944013.

[19] P. Routray, C. Kumar Swain, and S. Praya Mishra, "A survey on sentiment analysis," *International Journal of Computer Applications*, vol. 76, pp. 1-8, 2013.

[20] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, pp. 1345-1359, 2010.

[21] A. Ruchika and G. Latika, "A hybrid approach for sentiment analysis using classification algorithm," *International Journal of Computer Science and Mobile Computing*, vol. 6, pp. 149-157, 2017.

[22] J. Jeevanandam and S. Koteeswaran, "Sentiment analysis: A survey of current research and techniques," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 3, pp. 3749-3757, 2015. Available at: https://doi.org/10.15680/ijircce.2015.0305002.

[23] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of NAACL-HLT*, 2016, pp. 1480-1489.

[24] R. L. Vieriu, A. K. Rajagopal, R. Subramanian, O. Lanz, E. Ricci, N. Sebe, and K. Ramakrishnan, "Boosting-based transfer learning for multi-view head-pose classification from surveillance videos," in *Proc. 20th Eur. Signal Process. Conf. (EUSIPCO), Aug*, 2012, pp. 649-653.

[25] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1746-1751.

[26] B. McCann, J. Bradbury, C. Xiong, and R. Socher, "Learned in translation: Contextualized word vectors," in *Advances in Neural Information Processing Systems*, 2017, pp. 6294-6305.

[27] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365, DBLP:journals/corr/abs-1802-05365, CoRR, volume = abs/1802.05365*, 2018.

[28] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun, "Very deep convolutional networks for text classification," *arXiv preprint arXiv:1606.01781, CoRR volume = abs/1606.01781*, 2016.

[29] R. Johnson and T. Zhang, "Supervised and semi-supervised text categorization using LSTM for region embeddings," *arXiv preprint arXiv:1602.02373*, 2016.

[30] A. M. Dai and Q. V. Le, "Semi-supervised sequence learning," in *Processing Advances in Neural Information Processing Systems*, 2015, pp. 3079-3087.

[31] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *arXiv preprint arXiv:1801.06146, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), Melbourne, Australia, Association for Computational Linguistic*, 2018, pp. 328–339.

[32] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011.

[33] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the Conference Empirical Methods Natural Language Process*, 2013, pp. 1631-642.

[34] J. Blitzer, M. Dredze, and F. Pereira, "Domain adaptation for sentiment classification," in *Proceedings of the 45th Annual Meeting of the Association Computational Linguistics*, 2007.

[35] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *Stanford, CA, USA, Tech. Rep. CS224N*, vol. 1, 2009.

Views and opinions expressed in this article are the views and opinions of the author(s), Review of Computer Engineering Research shall not be responsible or answerable for any loss, damage or liability etc. caused in relation to/arising out of the use of the content.