Systematic Survey on Sentiment Analysis

Shubham Kumar Jain
Computer Science and Engineering Department
National Institute of Technology Hamirpur
Hamirpur, India
shubhamkjain218@gmail.com

Pardeep Singh
Computer Science and Engineering Department
National Institute of Technology Hamirpur
Hamirpur, India
avagaman@gmail.com

***Abstract*—Sentiment Analysis and Opinion Mining have been of great interest to the researchers during recent years. It is the process of classifying the opinions or sentiments according to the polarity of the text into positive, neutral and negative. Most of the organizations and industries highly depend on data analytics for their planning and decision-making process. Opinion mining and sentiment analysis have great importance in our day-to-day decision making from purchasing products and services to making investments. In this survey, we briefly incorporated the approaches and techniques proposed by researchers in recent investigations along with the issue related to sentiment analysis and opinion mining.**

Keywords—Sentiment Analysis; Opinion Mining; Natural Language Processing; Lexicon Based Approaches; Machine Learning Based Approaches

I. INTRODUCTION

Due to high-speed internet and attractiveness towards the micro-blogging websites, social networking, and blogging websites, there is a very large amount of information data is present online in the form of text sentences at these platforms. This text information may be utilized for various purposes including scientific investigations from different perspectives whether political or social [1]. Apart from this, the product and service based industries who are willing to improve their offerings may consider the rich responses and seek the large benefits out of them [2], [3]. And the customers on the other way, who willing to purchase the products or services could take insights about them before purchasing by knowing their positive and negative points and buy smartly. Moreover, with the use of this online information, many applications can be developed such as movies rating applications by making use the user's reviews on the internet about the movies [4], which is not possible otherwise.

Users widely use social media websites such as Twitter, Facebook, LinkedIn, and YouTube etc. to share their views regarding various events, products, and services at almost every second from all over the world.

A large increment is detected in the popularity of sentiment analysis, we have collected data about search result corresponding to sentiment analysis from Google trends and plotted in Figure 1 that depicts massing increase in the interest of people in the field of sentiment analysis in recent years. According to [5], popular social media website, Twitter, recorded around 25 billion of Tweets within one year.

And based on the negative reviews that users found on the internet websites almost 80 percent of the customers

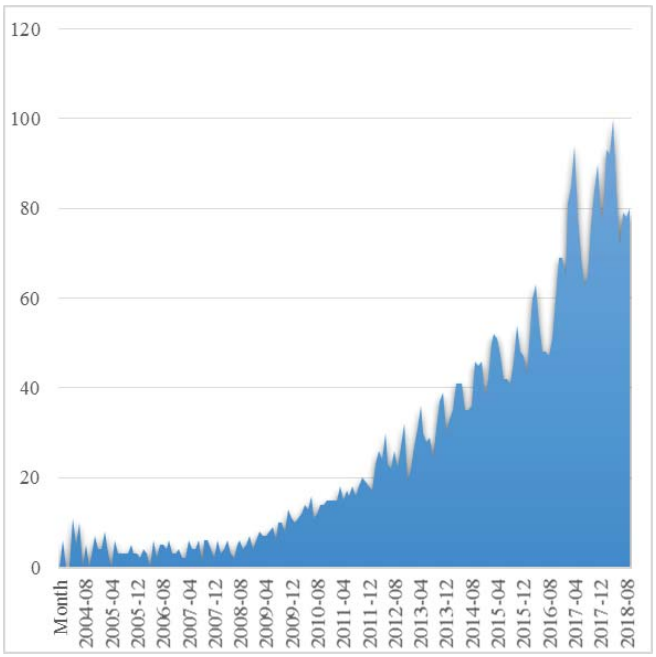

Fig. 1. Google Trends search results corresponding to the keyword "Sentiment Analysis".

changed their minds to purchase. This directly implies that businesses are greatly influenced by the comments of users on the internet about their products or services. Traditionally, Most of the organizations conduct their own well-organized surveys for maintaining the customer's relationships. These surveys may provide good estimations to the organizations but if they are organized on a very large volume, then it may be very expensive.

Sentiment Analysis is a method in which make use of computational models to categorize the views that are stated in the text sentences in order to find out whether the views about a specific topic, service or product, movie, etc. are neutral, positive or negative. It is a great application of Natural Language Processing techniques to extract the required information from the large source of text data.

Since information present in an online platform is not structured, so one of the aspects of the sentiment analysis is pre-processing such text-sentences and thereby categorize them based on their polarity like neutral, positive and negative. Another aspect is to determine whether the text sentence is subjective depicting the writer's views or opinions and objectives or stating the pure facts. Sentiment analysis is carried out on different levels such as phrase level, sentence level, and document level.

In case of document level analysis the complete document is assigned a single polarity class such as positive or negative

978-1-5386-6373-8/18/$31.00 ©2018 IEEE

while in sentence level, firstly the document is divided into the sentences and then these sentences are categorized into positive, negative or neutral whereas the phrase level sentiment analysis text is analyzed deeply and then aspects or phrases are identified and then these phrases are classified into negative, positive or neutral and it is also referred to as aspect-based analysis.

II. Related Work

There is vast growth in the research work regarding the opinion mining and sentiment analysis in the last decade. Previous work came up with the various methodologies and techniques relating to sentiment classification of the text data. This section includes some of the previous researches in the field of opinion mining and sentiment analysis.

P. Bhoir and S. Kolte in [9] implemented a rule-based system on movie review dataset with the help of the SentiWordNet approach and lexicon-based approach to facilitate user to analyze the movie on different aspects.

A. Salinca in [10] performed classification on business reviews from Yelp challenge dataset that consist of 15,585 business and 3,35,022 users' reviews [11] with the help of sentiment analysis using different feature extraction methods (Natural Language Toolkit and Word Sense Disambiguation) and classifiers out of that linear Support Vector Classifier and Stochastic Gradient Descent perform with accuracy of 94.4 % while Naïve Bayes and logistic regression perform slightly worst.

In other work, Seyed-Ali Bahrainian and Andreas Dengel in [8] proposed a new hybrid approach for polarity detection on Twitter data that consists of a preprocessor module for preprocessing the raw text data and a lexicon-based module for sentiment feature generation and a machine learning module containing linear support vector machine classifier. This hybrid approach resulted to be better than state-of-art methods available at that time with an overall accuracy of 89.13%.

Further Tripathi, P., Vishwakarma, S. and Lala, A. in [12] applied data mining techniques for classification of tweets and compared the performance of the two classifiers named Naïve Bayes and K-Nearest Neighbor resulting more accurate results in case of K-Nearest Neighbor, though the accuracy in this work result out to be relatively less as compared to other machine learning approach such as Linear Support Vector Machine.

Woldemariam, Y. in [15] used sentiment analysis in a cross-media analysis framework on Stanford Sentiment Treebank labeled using Amazon Turk with the help of Lexicon based approach with an overall accuracy of 38.45% and Recursive Neural Tensor Network with an overall accuracy of 48.34% and compared between the two broad categories of sentiment analysis (lexicon based and machine learning approach) by taking a sample methods from each and concludes that Recursive Neural Tensor Network method is more accurate for sentiment analysis on forum discussion posts.

Mumtaz, D. and Ahuja, B. in [17] applied Senti-lexicon algorithm which is a semi lexicon approach for sentiment analysis over movie reviews from Twitter with an overall accuracy of 70%. The algorithm is simple, versatile and feasible than other machine learning algorithms but the accuracy of this approach is lesser than other machine learning algorithms.

In [16] Khatri, S. and Srivastava, A. used Artificial Neural Network on the market data from Yahoo and Twitter between 01/01/2015 to 22/02/2015 to predict the stock market using the sentimental analysis by analyzing the moods of the users with an accuracy of 86.7%, although it would be better if data over a relatively large period taken in this work.

Rohini V, Merin Thomas and Latha .C.A in [14] performed sentiment analysis on regional language (Kannada) on movie reviews from Kannada websites with an overall accuracy of 79% by applying machine learning approach with Decision tree classifier and concludes that the direct analysis on regional language gave more accurate results than analysis of machine-translated English language though extracting knowledge from language other than English is more challenging task.

In [13] Linlin You and Bige Tuncer proposed Crowd-calibrated Geo-sentiment analysis mechanism and applied it on Instagram and Twitter feeds so as to facilitate the city planners and designers and the authorities to make a better decision on planning and designing the livable places. But it doesn't include multi-source data while deciding the better livable places like mobile network data, weather data, mobility data, place rating data due to which the overall accuracy came out to be 74.71%.

Rincy Jose and Varghese S Chooralil in [7] applied sentiment analysis to predict the election results. They used the Twitter streaming API tool for extracting the text data. And applied classifier ensemble approach for sentiment classification, that combines the outputs from a set of classifiers so as to minimize the risk of choosing inappropriate classifier and thus it resulted into the better performance to predict the election results on unseen data with an overall accuracy of 71.48%.

Mala, P. and Devi, S. in [18] analyzed the response towards a particular product on Facebook posts, with the help of users comments and reactions towards a post containing a product with an overall accuracy of 69.92%. They used graph API for extracting data and performed data analysis using Natural Language Toolkit API.

Further in [19] Samal, B., Behera, A. and Panda, M. compared among promising machine learning classifiers for sentiment analysis and figured out the best supervised machine learning algorithm for sentiment analysis by applying the seven promising supervised machine learning classifiers(which are Naïve Bayes, Multinomial Naïve Bayes, Bernoulli Naïve Bayes, Logistic Regression, Stochastic Gradient Descent, Linear Support Vector Machine(SVM)/ Linear Support Vector Classifier(SVC), Nu-SVM/ Nu-SVC) over movie reviews dataset and out of these Linear SVC/SVM yielded an accuracy of 100% over large data.

Zhang, X. and Yu, Q. in [22] proposed a solution to high dimensionality and high sparseness problem of the feature vectors. They used Word2Vec and ISODATA clustering algorithms to perform sentiment analysis on Hotel reviews and which provides high accuracy compared to other methods available but relatively high computation is required.

Park, C. and Seo, D. in [21] proposed a new criteria to decide better among the three popular artificial intelligence assistants that are Google Assistant, Cortana, and Siri by sentiment analysis of twitter corpus through lexicon named Valence Aware Dictionary and Sentiment Reasoner (VADER) along with Kruskal-Wallis test and Mann-Whitney test and calculated the respective ranks among which the Google Assistant stood first and Siri stood last. But they considered only English tweets in the experiment, the result may vary if the other languages were also considered.

In [20] Zvarevashe, K. and Olugbara, O. investigated the performance of different learning algorithms(Naïve Bayes Multinomial, Sequential minimal optimization, Complement Naïve Bayes Composite hypercubes on iterated random projections (CHIRP)) on Sentiment Polarity Based Model (SPBM) on OpinRank dataset and finds Naïve Bayes multinomial classification algorithm to be better with an accuracy of 80.9%. They used a dataset with unlabeled data that gives flexibility of customized experiment but it doesn't deal with the wrongly labeled data item, the model treats the item as neutral if the label not present in the dictionary.

III. Issues And Challenges

In the field of sentiment analysis and opinion mining, we have observed the following issues and challenges across the recent research work.

- One of the issues in sentiment analysis is that most of the work focuses only on English text data and thus in English, most of the resources (for example lexicons) are there. So it becomes difficult to apply this research to some other language text data. Therefore it is necessary to boost the study in other languages also.
- Another issue is to deal with the sarcastic text, that may be misinterpreted and incorrectly classified [23].
- As it is hard to collect the labeled data and it is expensive too, thus it is required that the classification of sentiments is done with the help of unlabeled data or insufficient labeled data which is still a challenging task [24].
- There are conditional sentiments that contain the actions that might occur in future also lead to incorrect polarity classification.
- The performance of the sentiment analysis also suffers from fake and spam reviews.
- Most of the works used the text data from social media websites, which suffers from poor spellings, abbreviations, poor grammar and punctuations that also lead to incorrect classification.
- The techniques that are built on supervised learning provides better results but needs knowledge base and the learning. According to, [25], a lot of time and effort is needed for creating this base. According to the authors lexicon based approached provides high accuracy but as there is lack of availability of lexicons in other languages it diminishes recall.

IV. Techniques And Approaches Used In Previous Works

In the recent investigations regarding sentiment analysis and opinion mining, the researchers have explored a variety of techniques in their work. This section summarizes the broad categories of the approaches and the techniques adopted in some of the previous work related to sentiment analysis and opinion mining. The very first step in the process of sentiment analysis is to choose among the datasets available and preprocessing that includes stemming, part of speech tagging, tokenization and removing the stop words and punctuations from the raw text data and then comes the task of sentiment classification after feature selection.

Figure 2 depicts the broad categories of approach to sentiment analysis or opinion mining. There are mainly three categories for classification of sentiments or opinions into negative, positive and neutral categories which are the lexicon-based approach, machine learning based approach, and hybrid approach combining both lexicon based and machine learning approaches.

A. Machine learning based approach

In machine learning approach the text features are extracted and used to perform classification. Due to its ability to handle a huge amount of data online and as it is automatic so more popular as compared to other techniques. The machine learning approaches mainly consist of a supervised learning approach, semi-supervised learning approach, and unsupervised learning approach.

1) Supervised learning based approach: A major part of machine learning techniques used for the purpose of sentiment classification is supervised machine learning techniques. They are widely used for classification purpose. In this, the data is divided into two sets one is a training set which is already labeled and other is test set data. The model is first trained on the training set to learn and then validated with the help of test set of data. The important algorithms in this category include Support Vector Machines, Naïve Bayes, Logistic Regression, Decision trees etc. Since these techniques are highly dependent on the training data so the accuracy may be influenced highly if there is a case of wrongly labeled data in training set or fewer data taken in training set.

2) Unsupervised learning based approach: Unsupervised learning techniques, unlike supervised learning approaches, don't make use labeled set or training set of data to train the model so these are useful when it is difficult to label the input data. But these require a very large amount of data to make the model learn otherwise it may produce incoherent results. These include algorithms like Word2Vector and K-means clustering as used for sentiment analysis on hotel reviews in [22].

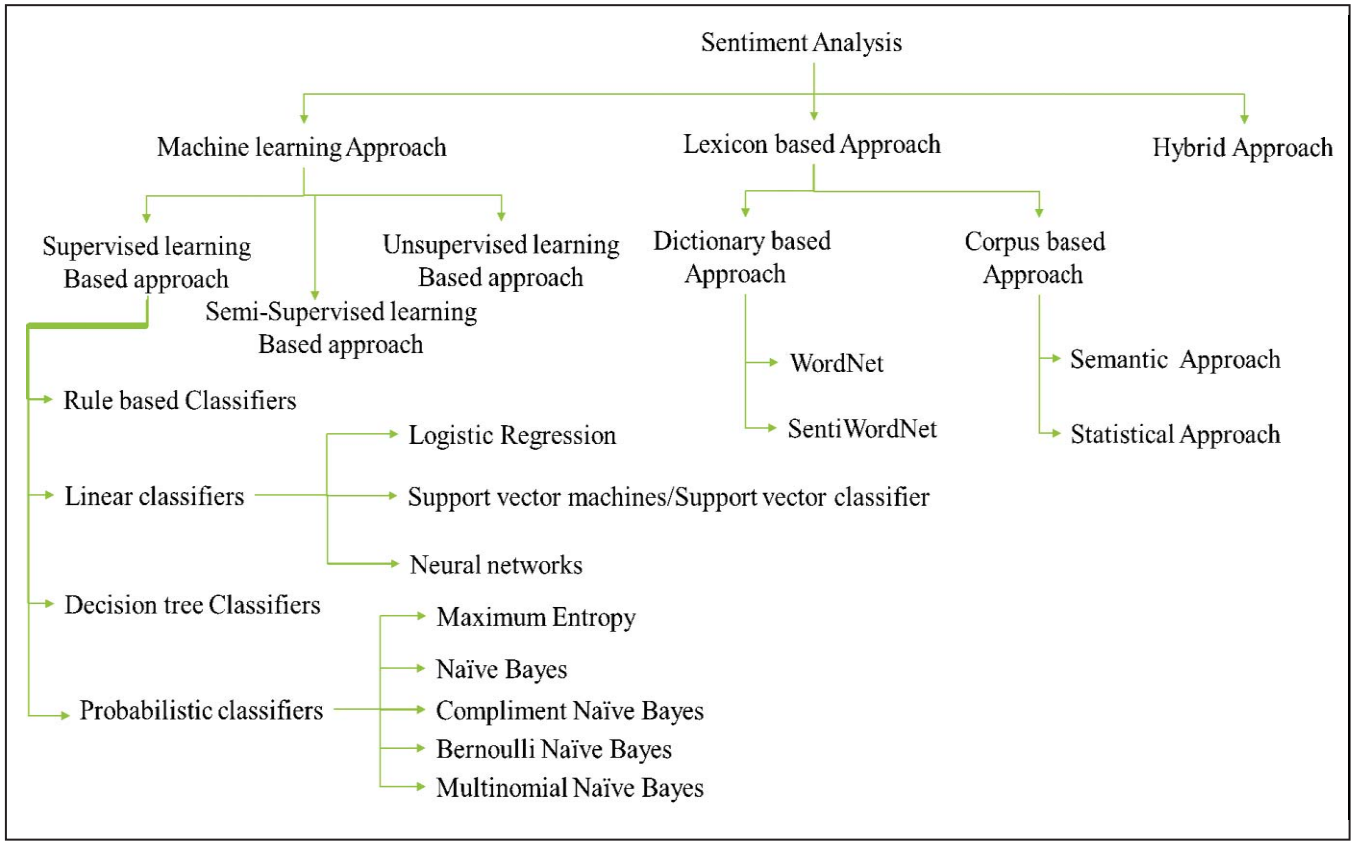

Fig. 2. Classification of approaches to sentiment analysis

3) Semi-supervised learning based approach: Semi-supervised learning approaches combines the advantages of both supervised and unsupervised learning approaches. In these, the model learns with the help of both labeled and unlabeled data [26]. These techniques were developed due to the lack of labeled data and to overcome the demerit of unsupervised learning approach as we are adding some of the earlier understanding to the unsupervised model [27].

B. Lexicon-based approach

Lexicon-based techniques mainly focus on determining the sentiment lexicon which is the collection of words in which every word contains a score that points to negative, neutral or positive nature of the text to be analyzed. The text information is analyzed with the help of this sentiment lexicon. For the given text information the scores for the subjective words are added separately and the maximum score decides the overall polarity [28].

The lexicon-based approach is mainly divided into two parts one is Dictionary based approach and other is the Corpus-Based approach.

1) Dictionary-based approach: This approach is based on determining the opinion seed from the text information and searching the dictionary for its antonyms and synonyms. Initially, a seed list is constructed by manually taking the opinion words which is difficult to get related opinion word due to limited context oriented text and thesaurus and dictionaries are then searched to figure out their antonyms and synonyms and later the synonyms are included into the list of seed words and the process is reiterated.

2) Corpus-based approach: Corpus-based approach deals with the constructing the list of seed opinion words and the list is expanded using the information from the corpus text. Corpus contains the pool of the text information mostly present on the specific domain so there is no problem of limited context oriented text information [29]. This can be achieved with the help of a semantic approach and statistical approaches.

C. Hybrid approach

The hybrid approach combines the method from both the lexicon-based approach and machine learning based approach so as to increase overall performance. In [31] Mukwazvure, A. and Supreethi, K. proposed a hybrid approach for the sentiment analysis of news comments in which they classified polarity by using sentiment lexicon and then they trained machine learning algorithms with the help of the outcomes from the lexicon based methods.

TABLE I. APPROACHES IN SENTIMENT ANALYSIS

S. No.	Approach	Authors
1	Lexicon based approach	[9], [28], [29]
2	Machine learning based approach	[7], [10], [13], [14], [16], [18], [20], [22], [27]
3	Hybrid approach	[6], [8], [31], [32], [33]

V. Conclusion

This paper presents an overall systematic survey of opinion mining and sentiment analysis. This paper analyzes the recent trends in sentiment analysis and the techniques with accuracy varying from 38.45% to 100% (over a large data in case of linear Support Vector Classifier) have been observed. The supervised machine learning approaches are widely used and apart from it the other approaches to sentiment analysis are the unsupervised machine learning approach, semi-supervised learning approach, lexicon-based approach and a hybrid approach that came out to be a fruitful solution as it combines the advantage of both lexicon based and machine learning based approach.

Further this paper figures out the various issues and challenges from the previous work related to sentiment analysis. Though there are a number of researches available on sentiment analysis, it is found that still there are some open challenges such as fake/spam reviews etc. that need to be addressed properly in future work.

References

[1] B. Liu, "Sentiment analysis and opinion mining," Synthesis Lectures on Human Language Technologies, vol. 5, pp. 1-167, 2012.

[2] M. Gamon, A. Aue, S. Corston-Oliver, and E. Ringger, "Pulse mining customer opinions from free text," Proceedings of the 6th international conference on Advances in Intelligent Data Analysis, p.121-132, September 08-10, 2005, Madrid, Spain.

[3] H. Tang, S. Tan, and X. Cheng, "A survey on sentiment detection of reviews," Expert Systems with Applications: An International Journal, v.36 n.7, p.10760-10773, September 2009.

[4] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," In Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10 (EMNLP '02), Vol. 10. Association for Computational Linguistics, Stroudsburg, PA, USA, 79-86, 2002.

[5] "Meaningful Growth," Blog.twitter.com, 2018. [Online]. Available: https://blog.twitter.com/official/en_us/a/2010/meaningful-growth.html. [Accessed: 03- Oct- 2018].

[6] F. Bravo-Marquez, M. Mendoza, and B. Poblete, "Meta-level sentiment models for big social data analysis," Knowledge-Based Systems, vol. 69, pp. 86-99, 10// 2014.

[7] R. Jose and V. Chooralil, "Prediction of election result by enhanced sentiment analysis on twitter data using classifier ensemble approach," 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE).

[8] S. Bahrainian and A. Dengel, "Sentiment analysis using sentiment features," 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT).

[9] P. Bhoir and S. Kolte, "Sentiment analysis of movie reviews using lexicon approach," 2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC).

[10] A. Salinca, "Business reviews classification using sentiment analysis," 2015 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC).

[11] [Online]. Available: https://www.yelp.com/dataset_challenge/dataset.

[12] P. Tripathi, S. Vishwakarma, and A. Lala, "Sentiment analysis of English tweets using rapid miner," 2015 International Conference on Computational Intelligence and Communication Networks (CICN).

[13] L. You and B. Tuncer, "Exploring public sentiments for livable places based on a crowd-calibrated sentiment analysis mechanism," 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM).

[14] V. Rohini, M. Thomas, and C. Latha, "Domain based sentiment analysis in regional language-Kannada using machine learning algorithm," 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT).

[15] Y. Woldemariam, "Sentiment analysis in a cross-media analysis framework," 2016 IEEE International Conference on Big Data Analysis (ICBDA).

[16] S. Khatri and A. Srivastava, "Using sentimental analysis in prediction of stock market investment," 2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO).

[17] D. Mumtaz and B. Ahuja, "Sentiment analysis of movie review data using senti-lexicon algorithm," 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT).

[18] P. Mala and S. Devi, "Product response analytics in Facebook," 2017 International Conference on Intelligent Computing and Control Systems (ICICCS).

[19] B. Samal, A. Behera, and M. Panda, "Performance analysis of supervised machine learning techniques for sentiment analysis," 2017 Third International Conference on Sensing, Signal Processing and Security (ICSSS).

[20] K. Zvarevashe and O. Olugbara, "A framework for sentiment analysis with opinion mining of hotel reviews," 2018 Conference on Information Communications Technology and Society (ICTAS).

[21] C. Park and D. Seo, "Sentiment analysis of Twitter corpus related to artificial intelligence assistants," 2018 5th International Conference on Industrial Engineering and Applications (ICIEA).

[22] X. Zhang and Q. Yu, "Hotel reviews sentiment analysis based on word vector clustering," 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA).

[23] E. Cambria, C. Havasi, and A. Hussain, "SenticNet 2: a semantic and affective resource for opinion mining and sentiment analysis," in FLAIRS Conference, 2012, pp. 202-207.

[24] S. Zhou, Q. Chen, and X. Wang, "Active deep learning method for semisupervised sentiment classification," Neurocomputing, vol. 120, pp. 536-546, 2013.

[25] S. Bhuta and U. Doshi, "A review of techniques for sentiment analysis of Twitter data," In Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014 International Conference on, pages 583–591. IEEE, 2014.

[26] O. Chapelle, B. Schölkopf, and A. Zien, "Semi-supervised learning," Cambridge, Mass.: MIT Press, 2006.

[27] Z. Chen, A. Mukherjee, and B. Liu, "Aspect extraction with automated prior knowledge learning," in Proceedings of ACL, 2014, pp. 347-358.

[28] K. Hanhoon, Y. S. Joon, and H. Dongil., "Senti-lexicon and improved naı¨ve bayes algorithms for sentiment analysis of restaurant reviews," Expert Syst Appl ,39:6000–10, 2012.

[29] K. Fazel and I. Diana., "A bootstrapping method for extracting paraphrases of emotion expressions from texts," Comput Intell; vol. 0, 2012.

[30] A. Alnawas and N. Arıcı, "The Corpus-based approach to sentiment analysis in modern standard Arabic and Arabic dialects: A Literature Review," Journal of Polytechnic.

[31] A. Mukwazvure and K. P Supreethi, "A hybrid approach to sentiment analysis of news comments," 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions).

[32] R. Feldman, "Techniques and applications for sentiment analysis," Communications of the ACM, vol. 56, pp. 82-89, 2013.

[33] H. H. Lek and D. C. C. Poo, "Aspect-based twitter sentiment classification," in Tools with Artificial Intelligence (ICTAI), 2013 IEEE 25th International Conference on, 2013, pp. 366-373.