An overview of Sentiment Analysis Approaches

Shamsuddeen Hassan Muhammad
Department of Software Engineering
Bayero University , Kano
Kano, Nigeria
shmuhammad.csc@buk.edu.ng

***Abstract*—Sentiment analysis is a relatively new field of study at the intersection of computer science and linguistics that aims to find an opinion expressed in a text. It has received a swell of interest in both academia and industry. This paper provides an overview of the basic approaches for sentiment analysis task: machine learning-based approach and lexicon-based approach. The machine learning approach is based on training models on corpora annotated with polarity information and the lexicon-based approach is based on using sentiment lexicon. Recently, a hybrid approach is employed to leverage the strength of both two approaches**

***Index Terms*—sentiment analysis, opinion mining, lexicon-based, domain-specific, machine learning**

I. Introduction

The exponential growth of social media and dynamic websites has changed how people use an Internet from read-only participation to read-write participation. Participants now actively contribute their opinion rather than reading the Web content passively [1].This digital revolution and paradigm shift allows users to express their opinion and sentiment on many areas such as government, commerce, politics, education, health and many entities [2]. For example, on average, 6,000 tweets per-second are made on Twitter and 91.8 million blog posts are published every month on Wordpress only [2]. These raise the need to find user's sentiments through such a medium.

Traditionally, people, businesses and government use approaches such as survey to find feedback or opinion on a particular subject. For example, if people want to buy a new mobile device, they consult their friends, relatives or acquaintance who had bought a similar product or service for an opinion and recommendation which can be positive, negative or neutral.They use the feedback received to determine the worthiness of the product to prevent disappointment. However, single or few opinions may be biased. In the same way, businesses conduct survey and opinion poll to find users' opinion on their product with a view to improve customer services and marketing strategy. Also, government uses a survey to find people reaction and acceptance towards new and existing policies. However, with the rapid increase of user-generated and opinionated text, the classical tools such as survey and traditional Natural Language Processing techniques(NLP) for analysing and understanding users sentiment or opinion are sub-optimal [3] [4]. To this end, an efficient way of finding user sentiment from text is needed [5].

Sentiment analysis (SA) is a study that aims to find sentiment, opinion, emotion, attitude computationally from written text [6]. It is an offshoot of natural language processing. Depending on the domain ,it is often referred to with a different nomenclature such as: *opinion mining, opinion analysis, opinion extraction, sentiment mining, sentiment extraction, subjectivity analysis, emotion analysis, review mining* and many more terms are evolving. Two most widely used names that appear in the academic are sentiment analysis and opinion mining. In contrast, only the term sentiment analysis is widely used in industry [7].

Early research on textual information processing focused on mining and retrieval of factual information, such as information retrieval, text classification or text clustering. Research in sentiment analysis started relatively in the year 2001 [13] [8] and the phrase *”opinion mining”* was first use in 2003 [14]. In [15], they reported that 99% of all the research on sentiment analysis have been published after the year 2004. Fig. 1 highlights most prominent areas of research in the area of sentiment analysis.

Pang and Lee [8] identify three factors which triggered interest in sentiment analysis research. First, the rise of machine learning methods in natural language processing and information retrieval; Second, the availability of datasets for machine learning algorithms to be trained on. Thirdly, the realization of the fascinating intellectual challenges and intelligence applications that the area offers.

Sentiment analysis has been successfully applied in many domain and applications such as recommender systems, user reviews and politics [8].Businesses also use sentiment analysis to find consumer opinion on product and services to improve their business and service delivery [9].

There are two basic approaches for sentiment analysis; lexicon-based approach and machine Learning-based approach.The machine learning approach is based on training models with corpora annotated with polarity information. The Lexicon-based approach is based on using polarity of lexicons and it provides better accuracy [10]. Recently, a hybrid approach has been proposed and it exploit the strength of two or more techniques to offer better performance [11].

This paper aims to provides an overview of the sentiment analysis approaches. Section II presents different levels of sentiment analysis. Section III discusses the three approaches of sentiment analysis. Finally, section IV concludes the paper.

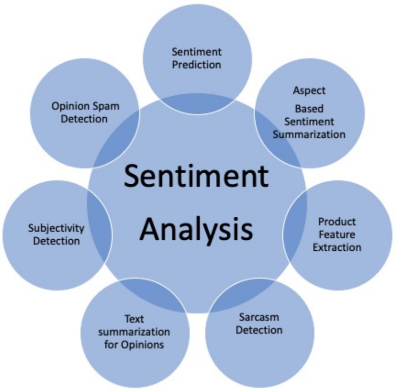

Fig. 1. Areas of research in sentiment analysis

II. Levels of Sentiment Analysis

According to [6], based on the levels of granularity, the sentiment analysis has been investigated at three different levels: document level, sentence level and aspect level. In contrast, Kumar & Sebastian [17] reported four different levels, with the addition of word level as depicted in Fig. 2.

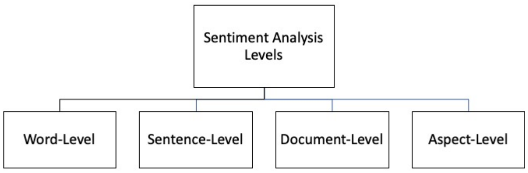

Fig. 2. Four Levels of sentiment analysis

1) Word Level: Sentiment analysis at word level involves finding adjective as a source of sentiment indicator. In the same way, other part-of-speech such as a noun, verb and adverb sometimes indicates subjectivity and opinion [16].

2) Sentence level: Sentiment analysis at sentence level deals with an opinion expressed in each sentence within a document. It finds the polarity of each sentence as positive, negative or neutral. Subjectivity classification is closely related to this concept; it deals with categorising sentences as either subjective sentences or objective sentences. However, subjectivity is different from sentiment because some objective statement may imply opinions e.g. "I bought a new computer yesterday and the battery does not last long" [6].

3) Document Level: In document level, the whole contents of the document are summarized to a single opinion. Therefore, it is assumed that each document contains an opinion on a single entity, thus, sentiment analysis at this level is not practicable for a document that contains sentiment on multiple entities [6].

4) Aspect level: This level of analysis is more difficult than sentence level and document level analysis. It is sometimes called feature level or feature-based opinion mining and summarization [17]. This level identified that any opinion without targets is meaningless and each opinion contains a target and sentiment. Therefore, the aim of the aspect level is to find sentiments on entities and/or their aspect. For example, the sentence *"The iPhones call quality is good, but its battery life is short "* evaluates two aspects, call quality and battery life of iPhone (entity). The sentiment on iPhones call quality is positive, but the sentiment on its battery life is negative. The call quality and battery life of iPhone are the opinion targets [6].

III. Approaches to Sentiment Analysis

There are two basic approaches to perform sentiment analysis. Machine learning-based and lexicon-based approach as shown in Fig. 3. Recently, hybrid approach of sentiment analysis has been explored to leverage the advantages of both machine learning and lexicon-base approaches [3] [11].

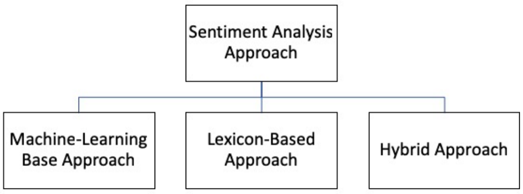

Fig. 3. Three approaches two sentiment analysis

A. Machine Learning Approach for Sentiment Analysis

Considerable research in sentiment analysis uses machine learning approach to perform sentiment classification. We briefly expound both the two approaches of supervised and unsupervised machine learning methods.

1) Supervised sentiment classification : Text classification has long been an existing research field and the task of sentiment classification is similar to the text classification. Text classification classifies text base on topics such as politics, religion and sports while sentiment classification classifies text to categories (classes) such as excellent, good neutral, bad and very bad. Some studies use numeric sentiment polarity values [7].

Similar to text classification method, supervised sentiment classification method uses a learning algorithm trained with sentiment-labelled data to classify an unseen document. The typical process of sentiment classification is shown in Fig. 4. First, standard text pre-processing, feature engineering and vector-space representation are applied to the training and test documents drawn from a problem domain. After that, a machine learning algorithm is employed to learn prediction model during a training phase. The model is then used in the testing phase to do classification (or regression) of unseen documents. One of the most important steps in the sentiment classification process outline is feature engineering. Feature engineering uses existing knowledge from the problem domain and create features that make machine learning more effective.The feature engineering process involves three phases of activates: feature discovery, feature selection and feature weighting.

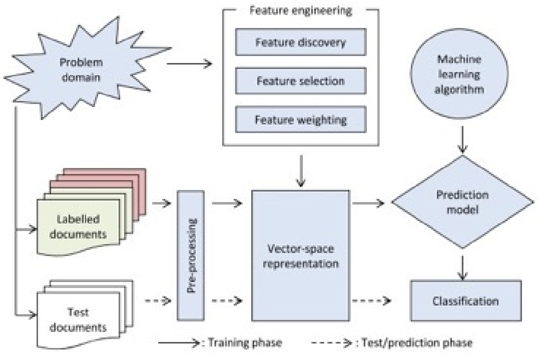

Fig. 4. Supervised sentiment classification method [33]

Previous research on sentiment analysis focus on using the standard machine learning algorithms such as *Naïve Bayes, Maximum Entropy and Support Vector Machine*. One of the pioneer study that experiments the three techniques performance on sentiment analysis task [13] reported that standard machine learning techniques perform better than human-produced baselines. However, the three machine learning methods performed poorly on sentiment classification compare to traditional topic-based categorization. The sub-optimal performance indicates that sentiment classification task is more difficult than topic classification. This is because topic can be easily identify by keyword alone while sentiment can be expressed in a subtler manner. For instance, *How could anyone sit through this movie?* contains no single word that is obviously negative. Hence fine-grain analysis is required with sentiment classification.

Recently, dedicated supervised methods for sentiment classification has been developed to improve accuracy. One of the techniques use score function [14]. The approach starts by training a classifier using a corpus of self-tagged reviews drawn from websites. Thereafter, the same corpus is then employed to improve their classifier before applying it to sentences obtain from web searches. Authors experimental result accuracy outperforms the traditional machine learning algorithms approaches.

2) Unsupervised sentiment classification: The unsupervised sentiment classification process is shown in Fig. 5. At the training phase, unlabelled documents are pre-processed and probabilistic topic modelling methods are employed to detect both topic and sentiment. Prior knowledge in the form of seed sentiment-bearing terms is required to guide the process. Consequently, the sentiment class of a text document can be determined based on the topic used to compose the document. Standard topic modelling approaches assume a three-layered hierarchical framework, where topics are associated with documents, and words are associated with topics. For sentiment detection, this framework is extended with an additional sentiment layer in between documents and topics or with sentiment classes as an additional topic model [26].

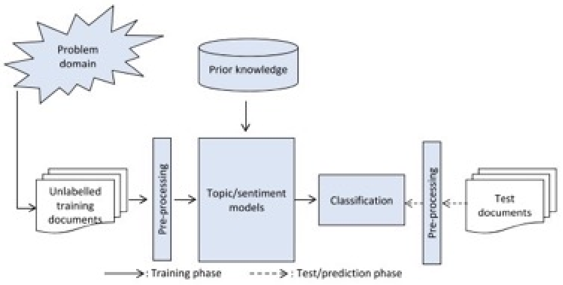

Fig. 5. Unsupervised sentiment classification method [33]

One of the pioneering unsupervised learning methods was proposed by Turney [13]. It is a simple unsupervised learning algorithm that classifies reviews base on the average semantic orientation of the phrases that contain adjectives and verbs. It classifies review as recommended if positive and not recommended if negative. The accuracy for machine learning-based approaches for sentiment analysis is not up to the mark compare to the lexicon-based approach [4].

B. Lexicon-based approach (Linguistic Approach)

The lexicon-based approach is based on using sentiment lexicon generated from either corpus or dictionary. It is sometimes called corpus-based approach if it uses lexicons generated from corpus or dictionary-based approach if it uses lexicons generated from a dictionary. It is workflow is shown in Fig. 6. The first step is the creation of sentiment lexicon or adoption of an existing one (which is mostly the case by many researchers). The next step is to pre-process the document to be classified and each word in the document is assigned the corresponding prior polarity from the sentiment lexicon. Finally, the prior polarities are adjusted to reflect contextual polarities (contextual analysis) and sum-up to find the sentiment orientation of the document. The sentiment orientation of the document is classified as either positive if the sum is positive or negative if the sum is negative and neutral if the final sum is 0. Variation of this exist and the difference is mainly based on what value is assigned to sentiment words in sentiment lexicon, how negation is handled etc., with the rapid increase of automatic generation of domain-specific lexicon, the lexicon-based approach is now leverage to provide better accuracy [27].

1) Sentiment Lexicon: Sentiment lexicon (lexical resource) is a dictionary of a lexical item with corresponding semantic orientation. It plays a significant role in sentiment analysis task. The lexical item conveys a single meaning and it can be words (e.g. good and bad), word senses, phrases (I am over the moon, it arrived) and idiomatic expression. The semantic orientation can be in several forms such as words (positive, negative or neutral) and phrases (strongly positive, mildly positive and strongly negative). A specific range of values is also used to indicate a ranking of the sentiment strength. For example, using 1 to 5 ranking with 1 has least ranking strength and 5 has the highest ranking strength. In this scenario, 3 being the middle is considered neutral [18]. Also, the accuracy

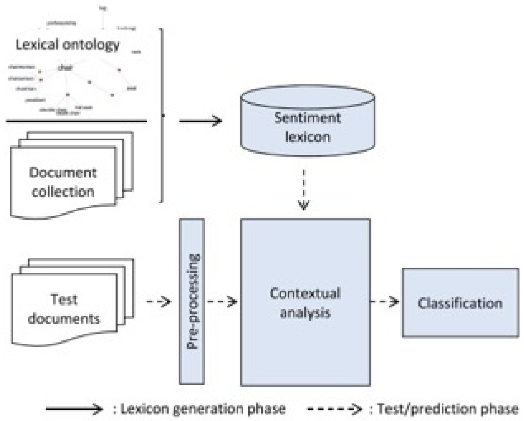

Fig. 6. Lexicon-based method for sentiment analysis [33]

of lexicon-base approach depends on the type of sentiment lexicons use.

a) General Purpose Lexicon: These are lexicons develop and use in sentiment analysis without any relation between the domain in question and the lexicon; they are nonspecific and can be used to find an opinion across domains such as a movie, social media, health and government. However, with the advantage of wide coverage, the lexicon losses accuracy because sentiment words are often domain-dependent [28]. A single word in one domain may contain positive polarity and contain negative polarity in another domain. Hence, the choice of word orientation is define by the domain in which the sentiment word is used. For example, the word *"suck"* appears to have negative and positive orientation in the following sentences: *"the camera we bought sucks,"*, the word *"suck"* have negative orientation in this sentence, but it can be used with positive orientation as in *"The vacuum cleaner we bought really sucks"*.

To exacerbate this problem, sentiment lexicons may have different sentiment orientation within the same domain. This makes the task of sentiment analysis even more difficult. For example, in camera domain, the word "long" have a different orientation in the following sentences. *"The battery life is long and It takes a long time to focus"*. The first sentence indicates positive opinion while the latter indicates negative opinion [24]. Consequently, several researchers use domain-specific lexicon for better accuracy.

b) Domain-Specific Lexicon: Due to the increase of application of sentiment analysis in several domains couple with need for accuracy , domain-specific lexicons are leveraged. They are generated from general purpose lexicon or a new one is generated from scratch. They increase the accuracy of sentiment analysis task. However, manual generation of the domain-specific lexicon for each domain is a laborious and mind-numbing task. To this effect, automatic methods for generation of domain-specific has been explored [12], [31], [32].

2) Sentiment Lexicon Generation Methods:

a) Manual Generation of Sentiment Lexicon: This approach consists of using an existing dictionary or corpus and manually select lexical items that have sentiment orientation. Thereafter, the lexical items are annotated manually with predefined sentiment strength.For example, 5-class sentiment strength is (-2 to +2). Because humans rather than machine annotate each lexical item , this method provides accuracy. Hence, sentiment analysis with manually generated lexicon achieve better performance. However, the method has drawback. It is time-consuming and daunting due to the inherent nature of manual activity involve. Also, it is limited to small coverage compared to automatically generated lexicon [18].

b) Automatic Generation of Sentiment Lexicon: In this approach, sentiment lexicon is generated automatically , therefore it eliminates time spend in manual approach. One of the popular automatic approach uses a seed word from which other sentiment words are generated automatically. Bootstrap approach is also employ and automatically ranks words based on a similarity measures [19]

3) Examples of Sentiment Lexicon: Some of the widely adopted sentiment lexicons are briefly explain in this section.

a) WordNet: This is an online English General lexical resource database. It contains adjectives, nouns and verbs group into synonyms set through semantic relation [20].

b) SentiWordNet: SentiWordNet is a lexical resource with a high level of coverage developed by Esuli and Sebastiani [21]. Positive, negative and neutral are three sentiment orientation used for each synset. It was developed from the *WordNet*. In SentiWordNet, words may contain different meaning and therefore different polarity. For example, "cold" may mean having a low temperature as in cold beer or without human warmth or emotion as in cold person. SentiWord uses glosses for each word entry to distinguish one from another [21].

c) WordNet-Affect: WordNet-Affect lexicon was created originally from WordNet synsets [22]. It consists of "Affective Knowledge" which describes moods, feelings and attitude. WordNet-Affect is one of the widely use lexicons because it is not limited to single-word concepts.

d) SenticNet: SenticNet is publicly available lexical resource for concept-level sentiment analysis. The lexicon includes both semantic and affective lexical unit. It provides over 30,000 multi-word expressions to enable fine-grain analysis of natural language opinion. It uses sentiment orientation between -1 and 1(-1 being extremely negative and +1 extremely positive) [23].

e) General Inquirer: This is a General Lexical system developed at Harvard for content analysis research in the behavioural sciences. The system uses two dictionaries: psycho-sociological dictionary and an anthropological dictionary used for studying themes in the folktales of many traditions and culture. The two dictionaries contain category of words. During sentence analysis, General Inquirer look-up these dictionaries and find in which category, the word belongs if it exists [24].

f) Bing Lius Opinion Lexicon: This is freely available sentiment lexicon developed by Bin Liu. It consists of English opinion lexicon being developed continuously. The lexicon contains a list of positive and negative words close to 6800 [25].

4) Domain Specific Lexicon-based Sentiment Analysis: The lexicon-based sentiment analysis approach performance reaches an optimal expectation when domain-specific lexicon is leverage. On the other hand, it gives sub-optimal performance when general purpose lexicon is leverage. To this effect, many research work on lexicon-based approach leverage domain-specific lexicons for better accuracy [12], [31], [32]. The approach is sometimes called: *Domain-dependent, Context-dependent, Domain-Oriented, Domain-based or Target specific lexicon-based approach*.

In [12], domain-specific lexicon from movie review corpora is automatically created and experimental results performance shows improvement over the manually created sentiment lexicons. Their method involves two steps, they generate corpus-based lexicon and each sentiment-bearing word is labelled with both positive or negative and polarity weight. Secondly, the lexicon is used in sentiment classification which shows improvement. Their approach is domain agnostic , therefore, very useful in creating domain-specific lexicons in many domains.

In the same way [29] proposed an approach for generating domain-specific lexicon through double propagation. Firstly, the technique uses a seed word to extract sentiment-bearing words and features. The extracted lexical items are then used iteratively to find new sentiment word and features until sentiment-bearing words are exhausted. They also proposed a method that assign polarity level to the sentiment words identified during sentiment extraction. Both proposed approaches provide satisfactory performance.

The study [11] devised a novel approach that exploits the idea of context coherency to automatically build a domain-focused lexicon for sentiment analysis. The context coherency is a phenomenon which explain that words with same polarity seems to always appear adjacent within context. The reported accuracy of this approach is 94% and proved to be effective and can be easily adopted in a different domain

A recent study [30] introduced a new domain-specific generation method from unlabelled review data. This approach is divided into two part, the first task is labelling the training reviews with polar values (negative and positive) and lexical unit with the higher ranking score are selected and used as training data. The second task uses the selected training data to obtain new domain-focused sentiment lexicon. The approach offered better performance when compared with other domain-specific lexicon base approach that uses SentMI and SenProf lexicon

C. Hybrid Approach

Until recently, a hybrid approach for sentiment analysis has been explored by researchers. They combine the strength of sentiment analysis approach for optimal accuracy. In their work [28], they leverage the strength of rule-based classification and supervised learning . The combined approach achieved higher accuracy when experiments on movie reviews, product reviews and Myspace comments. In [3], they perform Twitter sentiment analysis with a combination of lexicon-based approach and trained classifier. They claimed their result performs better than state-of-the-art baseline.

IV. Conclusion

We discussed an overview of sentiment analysis and the approaches to performing it in this paper. It is a sub-field of natural language processing that finds an opinion on human written text. It has been employ in different areas such as business and government. At a basic level, there are two approaches to perform sentiment analysis. Machine learning-based approach and lexicon-based approach. Until recently, a hybrid approach has been explored that combine the strength of two or more methods. The lexicon-based approach performance has been shown to outperform machine-learning approach when domain specific lexicons are employed. But, creating the domain-specific lexicon is a tedious and boring task. Therefore, as a solution to a manual generation of sentiment lexicon, novel approaches for automatic construction of domain-specific lexicon methods has been explored from recent literature

References

[1] A. Darwish, K. L. J. O. A. I. I. technology, 2011, The impact of the new Web 2.0 technologies in communication, development, and revolutions of societies, Citeseer

[2] K leen 121 Amazing Social Media Statistics and Facts.

[3] Combining Lexicon- and Learning-based Approaches for Improved Performance and Convenience in Sentiment Classification, pp. 150, May 2015.

[4] A. C. Forte and P. B. Brazdil, Determining the Level of Clients Dissatisfaction from Their Commentaries, in Computational Processing of the Portuguese Language, vol. 9727, no. 2, Cham: Springer, Cham, 2016, pp. 7485.

[5] B. L.2010, Sentiment Analysis and Subjectivity., Handbook of Natural Language Processing.

[6] B. Liu, Sentiment Analysis. Cambridge: Cambridge University Press, 2015, pp. 1384.

[7] B. Liu and L. Zhang, A Survey of Opinion Mining and Sentiment Analysis, in Mining Text Data, no. 13, Boston, MA: Springer, Boston, MA, 2012, pp. 415463.

[8] B. Pang and L. Lee, Opinion Mining and Sentiment Analysis, FNT in Information Retrieval, vol. 2, no. 1, pp. 1135, Jul. 2008.

[9] A Practical Guide to Sentiment Analysis, pp. 1199, Apr. 2017.

[10] J. Smith, Contextual Lexicon-based Sentiment Analysis for Social Media, pp. 1147, May 2016.

[11] A Hybrid Approach to the Sentiment Analysis Problem at the Sentence Level, pp. 16, May 2018.

[12] S. Almatarneh and P. Gamallo, Automatic Construction of Domain-Specific Sentiment Lexicons for Polarity Classification, in Trends in Cyber-Physical Multi-Agent Systems. The PAAMS Collection - 15th International Conference, PAAMS 2017, vol. 619, no. 4, Cham: Springer, Cham, 2017, pp. 175182.

[13] B. Pang, L. Lee, and S. Vaithyanathan, Thumbs up? Sentiment classification using machine learning techniques, in Proceedings of the 2002 conference on Emperical Methods in Natural Language Processing, 2002, pp. 7986.

[14] K. Dave, S. Lawrence, and D. M. Pennock, Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews, in Word Journal Of The International Linguistic Association, 2003, vol. 17, no. 5, pp. 519528.

[15] M. V. Mntyl, D. Graziotin, and M. Kuutila, The evolution of sentiment analysisA review of research topics, venues, and top cited papers, Comput. Sci. Rev., vol. 27, pp. 1632, 2018.

[16] A. Kumar, T. S. I. J. O. I. Systems, 2012, Sentiment analysis: A perspective on its past, present and future, mecs-press.net.

[17] B. Liu, Sentiment Analysis and Opinion Mining, Synth. Lect. Hum. Lang. Technol., vol. 5, no. 1, pp. 1167, 2012.

[18] S. Ahire, A Survey of Sentiment Lexicons, 2000.

[19] C. Banea, R. Mihalcea, J. W. LREC, 2008, A Bootstrapping Method for Building Subjectivity Lexicons for Languages with Scarce Resources., digital.library.unt.edu.

[20] G. A. Miller, WordNet: a lexical database for English, Communications of the ACM, vol. 38, no. 11, pp. 3941, Nov. 1995.

[21] S. Baccianella, A. Esuli, and F. Sebastiani, SentiWordNet 3 . 0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining SentiWordNet, Analysis, pp. 112, 2010.

[22] C. Strapparava and A. Valitutti, WordNet-Affect: an affective extension of WordNet, Proc. 4th Int. Conf. Lang. Resour. Eval., 2004.

[23] SenticNet 3: A Common and Common-Sense Knowledge Base for Cognition-Driven Sentiment Analysis.

[24] M. S. Smith, D. M. Ogilvia, P. J. Stone, D. C. Dunphy, and J. J. Hartman, The General Inquirer: A Computer Approach to Content Analysis., American Sociological Review.

[25] M. Hu and B. Liu, Mining and summarizing customer reviews. New York, New York, USA: ACM, 2004, pp. 168177.

[26] E. Cambria, D. Das, S. Bandyopadhyay, and A. (Editors) Feraco, A Practical Guide to Sentiment Analysis (Socio-Affecting Computing 5). 2017.

[27] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, Lexicon-Based Methods for Sentiment Analysis, vol. 37, no. 2, pp. 267307, May 2011.

[28] R. Prabowo, M. T. J. O. Informetrics, 2009, Sentiment analysis: A combined approach, Elsevier

[29] G. Qiu, B. L. 0001, J. Bu, and C. Chen, Expanding Domain Sentiment Lexicon through Double Propagation., IJCAI, pp. 11991204, 2009.

[30] H. Han, J. Zhang, J. Yang, Y. Shen, and Y. Zhang, Generate domain-specific sentiment lexicon for review sentiment analysis, Multimedia Tools and Applications, vol. 77, no. 16, pp. 2126521280, Aug. 2018.

[31] H. Kanayama, T. N. P. O. T. 2. C. on, 2006, Fully automatic lexicon expansion for domain-oriented sentiment analysis, dl.acm.org.

[32] Y. Lu, M. Castellanos, U. Dayal, and C. Zhai, Automatic construction of a context-aware sentiment lexicon: an optimization approach. New York, New York, USA: ACM, 2011, pp. 347356.

[33] MUHAMMAD, A.B. 2016. Contextual lexicon-based sentiment analysis for social media. Robert Gordon University, PhD thesis.