2018 3rd International Electrical Engineering Conference (IEEC 2018)
Feb, 2018 at IEP Centre, Karachi, Pakistan

An Overview of Lexicon-Based Approach For Sentiment Analysis

Azeema Sadia[1], Fariha Khan[2] and Fatima Bashir[3]

[1,3] Department of Computer Science, Bahria University Karachi Campus, Pakistan
([1] azeemasadia.bukc@bahria.edu.pk, [3] fatimabashir.bukc@bahria.edu.pk)
[2] Department of Electrical Engineering, Bahria University Karachi Campus, Pakistan
([2] farihakhan.bukc@bahria.edu.pk)

Abstract: Sentiment Analysis is the extraction of thoughts, attitudes and subjectivity of script or text to identify polarity i.e. positive, negative or neutral. There are three methods available for sentiment analysis, supervised, lexicon-based and hybrid approach, where the supervised method supersedes in performance from lexicon-based method and hybrid is a combination of both. The performance of supervised method is extremely reliant on on the excellence and the size of exercise data while on the other hand several lexical objects seem positive in the script of a domain while appearing negative at the same time in another domain therefore lexicon based analysis doesn't have high accuracy yet and optimizing it is still a very interesting research topic in the domain of Sentiment Analysis. This paper provides a comprehensive overview of the last updates in this field of lexicon based sentiment analysis along with their limitations and also shows our own methods' comparison of results for binary class classification and multiclass classification in the continuation of our future work.

Keywords: Sentiment Analysis, Lexicon, Polarity, Opinion mining

I. INTRODUCTION

With the huge volumes of data pouring in from every domain of every field like engineering, medical, management sciences, social media and others, there is a constant need of automated systems to classify that data based on different aspects. Sentiments analysis (SA) falls into the category of computational linguistics where the aim is to decide the outlook of the author towards a particular topic and coarsely speaking different approaches may be the result of people's beliefs, desires and feelings etc. Sentiment investigation is drawing in critical enthusiasm from open and corporate associations trying to mine client audits and online networking content for client assumption and supposition towards their items and administrations.

Specialists are currently creating systems for different sorts of sentiment investigation. A fundamental sort of assumption examination is estimation order – classifying bits of content into positive and negative polarity. Specialists have examined sentiment classification at the report level and also sentence level and even content section level.

Different algorithms have been applied so far but still the bottleneck lies in achieving remarkable accuracy. The analysis and applied processes are successful in identifying the polarity (depending on words) of a sentence but not the context i.e. a sentence can include positive words but it does not necessarily means that the sentence is positive and that will confuse the classifier.

II. METHODS IN SENTIMENT ANALYSIS

There are three approaches for broadly categorizing sentiment analysis: (a) Machine Learning based algorithms, (b) Lexicon based approach and (c) Hybrid Approach as shown in Fig. 1.

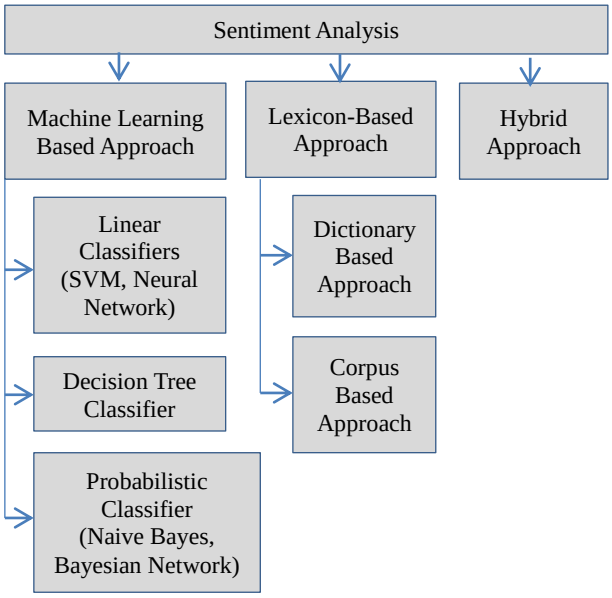

Fig.1 Sentiment Analysis Methods

A. Machine Learning Based Approach

Machine Learning based algorithms train the classifier from manually labeled data. However, the quality and coverage of training data have a high influence to performance of the classifier i.e. it requires a large database to be effective which is its only let down. This approach has better accuracy then lexicon-based.

B. Lexicon-Based Approach

This approach utilizes a sentiment lexicon to describe the polarity (positive, negative and neutral) of a textual content. This approach is more understandable and can be easily implemented in contrast to machine learning based algorithms. But the drawback is that it requires

the involvement of human beings in the process of text analysis.

The more prominent the information volume, the more noteworthy the test will be for sifting through the noise, identifying the sentiment and distinguishing helpful data from various content sources. Lexicon based approach can further be divided into two categories: Dictionary based approach (based on dictionary words i.e. WordNet or other entries) and Corpus based approach (using corpus data, can further be divided into Statistical and Semantic approaches).

C. Hybrid Approach

This approach is the amalgamation of both machine learning and lexicon-based methods.

This overview can be valuable for new comer scientists in this field as it includes a survey of different work on lexicon based sentiment analysis.

III. EMPIRICAL STUDY

Before the advancement of World Wide Web, people used to ask friends or family product recommendation but now the internet helps us to find out the experience of those who have used different products. In this era more individuals are sharing their outlooks with people through the internet. This huge data of people opinions on internet has started the trend to know about other's opinion. The year 2001 can be marked as the proper beginning of the research regarding sentiment analysis of people opinions. The term sentiment was first appeared in 2001 coined by Das et Al. [2] and Tong [3], Dave et al. published a research in the proceedings of the 2003 and used the word opinion mining the first time [4]. According to which the perfect opinion-mining implementation would be result generation for a particular product's attributes through searching and then categorizing the results in good, bad or mixed. Fig. 2 shows a typical process of Lexicon based sentiment analysis.

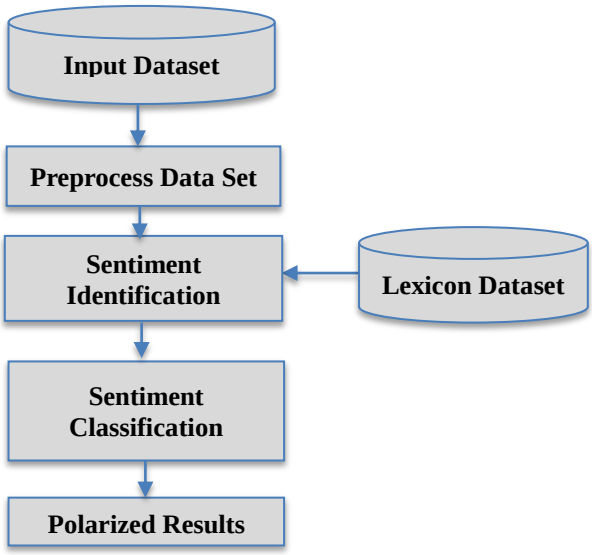

Fig. 2 General process of Lexicon based Sentiment Analysis

Sentiment analysis is a perplexing task. A few of the challenges are: Subjective part identification i.e. the part that contains the sentiments and deciding whether the word is subjective or objective is a difficult task. For example: A. "The customer's language was very crude". (Crude as opinion) B. "Crude oil is being imported". (Crude is objective). Dependence on domain is also of main importance as one word can be positive in one domain and negative in other. For example: The word "unpredictable" is positive in context of a movie but not in terms of amount of spices in dish. Detecting sarcasm and contextual meaning is also difficult as sarcasm means expressing negative comments in a unique way using positive words. For example: "The restaurant is too good when it comes to bill". Here the opinion is contextually negative but the words used are positive.

In the year of 2002 different text classification techniques were introduced like Naive bays, Support vector machine etc. A sophisticated solution was proposed focused on binary classification which handled the problem of model misfit apparent in some existing text categorization techniques [5]. Similarly in 2004, Hu, M. and Liu, B. produced a paper "Mining and Summarizing Customer Reviews", research was different from traditional work because they only mined those aspects of the product which the customer stated his opinions and identified if these opinions were negative or positive by using WordNet which helped them to distinguish the semantic orientation of opinion words [6].

Kanayama, H. and Nasukawa proposed the idea for clause level sentiment analysis in 2006. Their research titled as "Fully Automatic Lexicon Expansion for Domain-oriented Sentiment Analysis" described the methodology for clause level sentiment analysis in which they performed phrase restriction, the input document was separated into sentences. After that proposition detection the final step was polarity assignment which was assigned by comparing their lexicon polar item with the acknowledged propositions [7].

Another lexicon based method was given by Ding. X et al. in year of 2008, the research proposed a technique to identify the orientation of product reviews that were context dependent. Previous researches only considered unambiguous opinions articulated by adjectives and adverbs [8].

In 2011, Taboada M. et al. proposed a lexicon based approach for the sentiment analysis of the text. The idea proposed was Semantic Orientation Calculator that used thesauruses of words with their semantics. The purpose of SO-CAL was to assign the polarities (positive/negative) to the text. They also described the development of dictionary and used different dictionaries in order to know the performance of SO-COL [9].

Florian Wogenstein et al. in the year 2013 presented a paper in which sentence based opinion lexicon was used for the German language; they worked on the phrases from the insurance domain and analyzed the huge

difference in accuracy amid positive and negative statements [10]. In the same year Prabu Palanisamy et al. proposed Serendio taxonomy consisting of positive, negative, end words and expressions and also proposed their own sentiment calculation technique [11].

Alexander Hogenboom et al. in 2014 presented the idea for multi lingual sentiment analysis using lexicon based approach. Input text was translated into reference language. Sentiment scores were mapped to a new target sentiment lexicon from sentiment lexicon in the reference language, through traversing of associations amongst language-specific semantic lexicons [12]. In the same year Gaurangi Patil et al. described the data preprocessing and information retrieval using support vector machine. It was indicated that Support Vector Machine acknowledged particular properties of script for example High Dimensional feature space and sparse instance vector [13].

Sara Rosenthal et al. took part in Sem-Eval 2015 task 10: Sentiment Analysis on twitter. Input data set consisted of tweets about general topics. The collected tweets were largely tilted towards the neutral class that's why the imbalance of class was reduced by removing the tweets containing non sentiment words, for this purpose they used SentiWordNet as database of sentiment words. The degree of polarity was assigned the input tweets by using spontaneously created sentiment lexicons i.e. Hashtag Sentiment Lexicon and Sentiment140 Lexicon [14].

In year of 2016 new meta-level features for sentiment analysis was proposed. Three classifiers Vader, SentiStrength and SentiWordnet were used to guess the sentiment worth of every note. Precisely, the positive and negative sentiment scores of every note were extracted by proposed methods, along with joint and neutral scores specified by Vader [15]. Another research followed these steps to perform SA: the first step was preprocessing that's basically data cleansing in which noise in the data is removed by eliminating stop words and punctuation marks etc. Second step is the probability calculation of every term in a sentence seperately using unigram language model. The third step was to find sentiment of every word i.e. positive, negative and neutral which were calculated using a standard lexicon (National Research Council Canada (NRC) lexicons) as shown is Fig 3, 4.

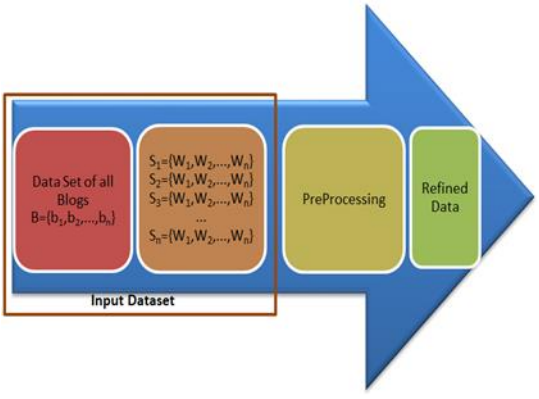

Fig. 3 Dataset Refinement Flow

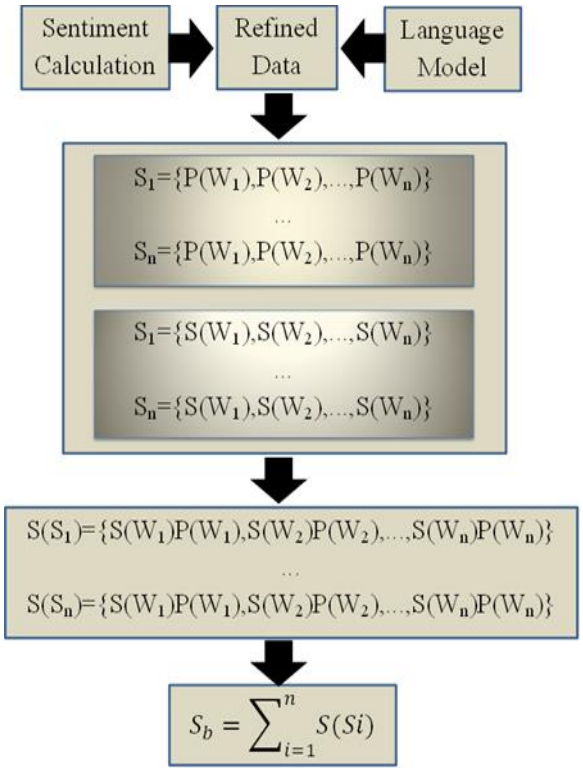

Fig. 4 Basic Block Diagram for SA Processes [16]

A comparison of different lexicons was performed in 2017 which introduced a lexicon named WKWSCI Sentiment Lexicon and compared it to five prevailing lexicons: Multi-perspective Question Answering (MPQA), National Research Council Canada (NRC), Hu & Liu Opinion Lexicon, Semantic Orientation Calculator (SO-CAL) lexicon, Subjectivity Lexicon, General Inquirer and Word-Sentiment Association Lexicon. The efficiency of the lexicons aimed at sentiment cataloguing at the sentence and document level was assessed by a news headlines dataset and an amazon product review data set. MPQA, Hu & Liu and WKWSCI, SO-CAL lexicons resulted in precision rates of 75%–77%. Hu & Liu obtained the highest accuracy with a naive method of totaling positives and negatives. The WKWSCI lexicon gained the precision of 69% [17].

Aung, K. Z. et al. used lexicon based approach to foresee teaching adequateness. A database English sentiment arguments was shaped as a lexical source to get the polarity of words [18]. This approach relied on bootstrapping using seed opinion words and online thesaurus. Mainly collection of set of views manually with recognized directions, and then to enhance this set by finding in the WordNet for substitutes and antonyms. The newly found words were included to seed list. The next cycle begins. The repetitive process loop stops if no different words were found. The semantic orientation score of joining words in all sentences are added to achieve the final polarity results [18].

The Table 1 below shows the summary along with technique limitation of the some papers that contributed in the area of lexicon-based sentiment analysis.

Table 1 Different Approaches used along with classification type, Data scope / Dataset and limitations

Year & Approach	Polarity	Data Scope & Dataset / source	Limitation
Year: 2002 Approach to Handling Model Misfit in Text Categorization [5]	Positive, Negative	**Data Scope:**1.Reuters-21578 dataset, 2.Usenet articles2 **Dataset / source:** Reuters newswire, Lang (1995)	Improved ways are required to find the performance of the base classifier during the training phase.
Year: 2002 Semantic Orientation for Unsupervised Classification of Reviews [19]	Positive, Negative	**Dataset / source:** Product reviews	Time limitation for queries, low level of accuracy for some application.
Year: 2004 Using Opinion Words [6]	Positive, Negative	**Data Scope:**Customer reviews **Dataset / source:** Amazon.com	The algorithm does not cater to pronoun resolution, defining the strength of opinions, and scrutinizing opinions expressed with adverbs, verbs and nouns.
Year: 2006 Automatic Lexicon Expansion for Domain- Oriented Sentiment Analysis [7]	Positive, Negative, Neutral	**Data Scope:** Japanese Reviews data set **Dataset / source:** Movie review data set - Turney, 2002, The human evaluation result - digital camera domain (Kanayama et al., 2004).	The approach is insensitive to deal with the complexity of human words during presenting their opinions about any product.
Year: 2008 Holistic Lexicon Based Approach [8]	Positive, Negative	**Data Scope:**Customer reviews **Dataset / source:** http://www.cs.uic.edu/~liub/FBS/FBS.htm/	The work is not able to find synonyms.
Year: 2011 Lexicon Based Approach [9]	Positive, Negative	**Data Scope:**Review text **Dataset / source:** 1. epinions.com 2. Texts from the Polarity Dataset (Pang and Lee 2004. 3. Text used in Bloom, Garg, and Argamon (2007).	This technique cannot analyze sarcasm.
Year: 2013 Simple and Practical Lexicon based Approach [11]	Positive, Negative, Neutral	**Data Scope:**Tweets **Dataset / source:** Twitter.com	Not suitable for word sense disambiguation like word good is identified as positive word but it can also be negative in sense when used as, "Good mile from here".
Year: 2013 Aspect Based Opinion Mining [10]	Positive, Negative	**Data Scope:**German phrases **Dataset / source:** http://nlp.stanford.edu/software/lex-parser.shtml	Incapable of dealing with verb-based phrases.
Year: 2013 Sentiment Analysis of Movie Reviews [20]	Positive, Negative	**Data Scope:** Movie Review Dataset **Dataset / source:** www.imdb.com	The only restriction is that it is domain specific and it is difficult to update the dictionary.
Year: 2014 Lexicon Based Approach [12]	Positive, Negative	**Data Scope:**Micro Blogs **Dataset / source:** 1.www.cs.york.ac.uk/semeval-2013/task2/ 2.https://dev.twitter.com	---
Year: 2014 Lexicon Based Approach [13]	Positive, Negative	**Data Scope:**German Phrases **Dataset / source:** http://www. teezir.com	Misinterpretation of text can cause the failure of the algorithm.
Year: 2015 Lexicon based approach [14]	Positive, Negative, Neutral	**Data Scope:**Tweets **Dataset / source:** dev.twitter.com	---
Year: 2016 Sentiment Based Meta-lexicon Based Approach [15]	Positive, Negative, Neutral	**Data Scope:**Short messages **Dataset / source:** 1. aisopos tw , 2. debate, 3. narr tw , 4. pappas ted, 5. pang movie, 6. sanders tw3, 7. ss bbc, 8. ss digg, 9. ss myspace, 10. ss rw,	---

		11. ss twitter, 12. Ss youtube, 13. stanford tw, 14. msemeval tw4, 15. vader amzn, 16. vader movie, 17. vader nyt, 18. vader tw, 19. yelp review	
Year: 2017 Comparative study [17]	Positive, Negative, Neutral	**Data Scope:** Amazon product review data set, news headlines data set.	---
Year: 2017 Lexicon-Based Approach for Students' Comments [18]	Positive, Negative, Neutral	**Dataset / source:** Department of Languages, the University of Computer Studies, Mandalay	This technique cannot analyze sarcasm.

VI. CONCLUSION

This survey paper presents an overview and recent updates in lexicon based sentiment analysis. The articles discussed explained the contributions to many sentiment analysis linked areas that use lexicon based analysis. After analyzing these articles, it is clear that the advancement in lexicon based sentiment analysis is still an open field for research. Most of the research is in English language but now the interest is increasing as there is a lack of resources and researches for other languages. WordNet is the most common lexicon sources. In almost all applications it is of utmost importance to consider the context of the text than just plain polarity and for that we still need enhancements in our algorithms.

IV. FUTURE WORK

In previous research work related to lexicon –based sentiment analysis on restaurant reviews unigram language model was incorporated with NRC lexicon status in order to achieve polarity score [16]. The polarity of the input was dependent on the result of unigram language model multiplied by the score of lexicon dictionary as shown in Fig.4 [16]. The Classification is generally of two types; the first is binary class i.e. dividing the reviews into two groups positive and negative, and multiclass i.e. dividing the reviews into more than two like in our case three groups positive, negative and neutral.

The results showed that 85.5% accuracy was achieved for binary class classification which decreased to 48% for multiclass classification because the inclusion of one more class i.e. neutral, increases the difficulty level as now the reviews have to be divided in three groups and differentiating between positive and neural, and negative and neutral becomes challenging. An example for positive review that can be considered as negative could be "The steak was nice it had killer flavor" here the word killer can mislead the sentiments because the algorithm is unable to identify the context of the reviewer. Another example of a neutral sentence that can be mistaken as positive can be "The ambiance is good and the food is ok".

We intend to use bigram and trigram to analyses whether the accuracy improves for multiclass classification or it affects the accuracy for binary class classification. We also aim to check and compare the accuracy of the proposed lexicon model using our own dataset with machine learning algorithms for future research.

REFERENCES

[1] Medhat, W., Hassan, A., & Korashy, H., "Sentiment Analysis Algorithms and Applications: A Survey," *Ain Shams Engineering Journal*, pp. 1093-1113, 2014.

[2] Sanjiv Das, Mike Chen, "Extracting Market Sentiment From Stock Message Boards," *Proceedings of the Asia Pacific Finance Association Annual Conference (APFA)*, 2001.

[3] Richard M. Tong, "An Operational System for Detecting and Tracking Opinions in On-Line Discussion," *Proceedings of the Workshop on Operational Text Classification (OTC)*, 2001.

[4] Dave, K., Lawrence, S., & Pennock, D. M., "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews," *Proceedings of the 12th International Conference on World Wide Web*, *ACM*, pp. 519-528, 2003.

[5] Wu, H., Phang, T. H., Liu, B., & Li, X., "A Refinement Approach To Handling Model Misfit In Text Categorization, *Proceedings Of The Eighth ACM SIGKDD International Conference On Knowledge Discovery And Data Mining*, *ACM*, pp. 207-216, 2002

[6] Hu, M., & Liu, B., "Mining And Summarizing Customer Reviews," *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM*, pp. 168-177, 2004.

[7] Kanayama, H., & Nasukawa, T., "Fully Automatic Lexicon Expansion For Domain-Oriented Sentiment Analysis, *Proceedings of the 2006 Conference On Empirical Methods In Natural Language Processing*, Association for Computational Linguistics, pp. 355-363, 2006.

[8] Ding, X., Liu, B., & Yu, P. S., "A Holistic Lexicon-Based Approach to Opinion Mining," *Proceedings of the 2008 International Conference On Web Search And Data Mining*, *ACM*, pp. 231-240, 2008.

[9] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M., "Lexicon-Based Methods for Sentiment Analysis," *Computational linguistics*, pp. 267-307,

2011.

[10] Wogenstein, F., Drescher, J., Reinel, D., Rill, S., & Scheidt, J., "Evaluation of an Algorithm for Aspect-Based Opinion Mining using a Lexicon-Based Approach," *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining, ACM*, pp. 5, 2013.

[11] Palanisamy, P., Yadav, V., Elchuri, H. Serendio, "Simple and Practical lexicon Based approach to Sentiment Analysis," *Proceedings of Second Joint Conference on Lexical and Computational Semantics*, pp. 543-548, 2013.

[12] Hogenboom, A., Heerschop, B., Frasincar, F., Kaymak, U., & de Jong, F., "Multi-lingual support for Lexicon-Based Sentiment Analysis Guided by Semantics," *Decision support systems*, pp. 43-53, 2014.

[13] Gaurangi Patil, Varsha Galande, Mr. Vedant Kekan, Ms. Kalpana Dange, "Sentiment Analysis Using Support Vector Machine," *International Journal of Innovative Research in Computer and Communication Engineering*, 2014.

[14] Sara Rosenthal, PreslavNakov, Svetlana Kiritchenko, Saif M Mohammad, Alan Ritter, VeselinStoyanoy, *" SemEval-2015 Task 10: Sentiment Analysis in Twitter,"* Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pp. 451–463, 2015.

[15] Canuto, S., Gonçalves, M. A., & Benevenuto, F., "Exploiting New Sentiment-Based Meta-Level Features for Effective Sentiment Analysis," *Proceedings of the ninth ACM international conference on web search and data mining*, pp. 53-62, 2016.

[16] Sadia, A., "Sentiment Analysis using Language Model," International Conference on Emerging Trends in Engineering, Sciences and Technology, pp. 50-53, 2016.

[17] Khoo, C. S., & Johnkhan, S. B., "Lexicon-Based Sentiment Analysis: Comparative Evaluation Of Six Sentiment Lexicons," *Journal of Information Science*, 2017.

[18] Aung, K. Z., & Myo, N. N., "Sentiment Analysis of Students' Comment using Lexicon Based Approach," *Computer and Information Science (ICIS), IEEE/ACIS 16th International Conference IEEE*, pp. 149-154, 2017.

[19] P. D. Turney, "Thumbs Up or Thumbs Down?: Semantic Orientation Applied To Unsupervised Classification Of Reviews," Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL), pp. 417–424, 2002.

[20] V. K. Singh, R. Piryani, A. Uddin, P. Waila, "Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification," International Multi-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), 2013.