# A study of Machine learning based, Lexicon based, and Hybrid sentiment analysis based methods

Fahd Saleh Alotaibi[1,*], Khaled Hamed Alyoubi[1,#]
[1]*Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia*
*(E-mail: [*]fsalotaibi@kau.edu.sa, [#]kalyoubi@kau.edu.sa)*

***Abstract*—**Sentiment analysis examines attitudes, emotions and sentiments of people concerning various products and services. Due to the vast amount of netizens' data available on the internet, manually analyzing data and making decisions has become challenging. Therefore, researchers are motivated to perform a wide variety of sentiment analyses. Sentiment analysis is a technique characterized as opinion mining for categorizing text into negative class, positive class, or neutral class. This paper discusses a study of Machine learning based, Lexicon based, and Hybrid sentiment analysis based methods.

***Keywords*—***Sentiment analysis; lexicon; machine learning;*

## I. INTRODUCTION

In present time, the growth of online social networking websites such as product reviews sites, micro-blogging sites, Twitter, Facebook and other social networks has prompted the advancement in sentiment analysis (SA). Due to the vast amount of netizens' data available on the internet, manually analyzing data and making decisions has become challenging [1]. Therefore, researchers are motivated to perform a wide variety of sentiment analyses. Sentiment analysis is a technique characterized as opinion mining for classifying text into negative class, positive class, or neutral class [2] [20]. Sentiment analysis based approaches are of lexicon oriented, machine learning oriented, deep learning oriented [22] [23] and hybrid. Various applications include product analysis, movie review analysis, stock prediction, government intelligence, recommender systems, etc.

## II. RELATED SURVEY

Related survey section, discusses existing research work studied by several researchers in sentiment analysis which leads to the development of many methodologies and models.

Hossen and Dev (2021) [6] used the SentiWordNet dictionary for performing lexicon-based analysis of sentiments using IMDB dataset for movie reviews. Kumar et al. (2019) [7] combined lexicon and machine Learning based features using IMDB dataset for movie review for performing sentiment analysis. Various features were used like TF (Term Frequency), TF-IDF (Term Frequency-Inverse Document Frequency), count of Positive words, Negative words, negative Connotation and positive Connotation which boosted the overall accuracy of model. RLPI (Regularized Locality Preserving Indexing) based feature selection technique was also used to minimize the dimensionality of feature data of machine learning features. Gopi et al.(2020) [21] used the improved RBF kernel of SVM to categorize the tweets. Author (2020) [5] performed sentiment analysis on Hindi and English tweets related to "Mann Ki Baat" using term frequency and various lexicon feature extraction approaches. Chauhan and Sutaria (2019) [19] experimented with semiotics for performing sentiment analysis. Here they have used text and semiotics together for calculating the score values of various tweets. Hemalathaand Ramathmika(2019) [8] suggested a machine learning oriented technique to find the sentiments of a yelp review dataset using POS tagging to extract adjectives from reviews. H. Darshan et al. (2019) [8] performed similar work on the IMDB movie review dataset, using the RLPI feature selection technique. Kumar et al. (2018) [9] used a weight assignment approach called Point wise Mutual Information (PMI) based on the SentiWordNet lexical resource to perform sentiment analysis on three datasets. Adewole et al. (2021) [10] suggested a feature selection approach that relies on a hybrid filter and wrapper method to improve sentiment analysis. Comparing results demonstrates the superiority of RF with a decrease in the number of features by 95% from the original dataset. Sunitha et al. (2019) [11] used a supervised classification algorithm to conduct sentiment analysis on Twitter tweets and a US airline dataset. The authors employed an advanced preprocessing method and TF-IDF feature vector extraction to enhance sentiment analysis classification accuracy.

Several researchers used hashtags as a feature in sentiment analysis. Alfina et al. (2017) [12] mainly focused on the characteristic of a hashtag count for sentiment analysis. S. Naz et al. (2018) [13] used an SVM classifier on the SemEval 2016 Twitter dataset to analyze the sentiment. To improve a traditional n-gram based classifier, they combined internal n-gram based score with an exterior sentiment based score. In [14], author Eng et al. (2021) also proposed a hybrid classification approach using lexical resources and machine learning classifiers. Here, the author used a variety of lexicons, including sentiment, negation, intensifier, emoticon, and so on, and combined with machine learning classifiers to calculate the sentiment score. Khoo and Johnkhan (2018) [15] examined the performance of WKWSCI, a unique lexicon-based technique, on Amazon product reviews and news headlines datasets. Apart from word features, various researchers used emoticons feature for sentiment analysis. John et al. (2019) [16] recommended hybrid lexicons followed by the handling of emoticons, capitalization, and repeated letter and discourse structure to increase the accuracy.

According to the literature review described in Table I we concluded that little research work has been done on the

linguistic aspects of hashtags, emojis, and emoticons for sentiment analysis. Furthermore, there has been limited study on combining lexical features with an ensemble of machine learning classifiers.

## III. FEATURES AND DATASET USED IN EXISTING METHODS

This section describes various features and datasets applied in various existing techniques (lexicon oriented techniques /Machine learning oriented methods/ Hybrid techniques) for analysis of sentiments as per Table I.

TABLE I. LITERATURE REVIEW OF EXISTING SENTIMENT ANALYSIS METHODS

| Author Name | Dataset | Feature selection | Method Used |
|---|---|---|---|
| Hossen and Dev (2021)[6] | IMDB movie review dataset | SentiWordNet dictionary | Lexicon based approach |
| Kumar et al. (2019) [7] | IMDB movie review dataset | TF and TF-IDF machine learning based features. Positive, Negative word count, positive and negative Connotation lexicon based features | Hybrid approach |
| Darshan et al. (2019) [8] | IMDB movie review dataset | RLPI feature selection technique | Machine learning based approach |
| Hemalatha and Ramathmika (2019) [8] | Yelp review dataset | POS tagging | Machine learning based approach |
| Kumar et al. (2018) [9] | IMDB movie review, Yelp review, Amazon dataset | PMI Weight assignment model based on SentiWordNet 3.0. | Hybrid approach |
| Adewole et al. (2021) [10] | IMDB movie review, Yelp review, Amazon, and Kaggle dataset | Hybrid filter and wrapper feature selection technique | Machine learning based approach |
| Sunitha et al. (2019) [11] | Twitter review and a US Airline Twitter Sentiment dataset | TF-IDF feature vector extraction | Machine learning based approach |
| Alfina et al. (2017) [12] | Manual dataset | Positive, negative word count, Positive and negative hashtag count | Hybrid approach |
| Naz et al. (2018) [13] | SemEval 2016 dataset | n-gram features | Machine learning based approach |
| Eng et al. (2021) [14] | Yelp review dataset | Sense disambiguatio | Hybrid approach |
| | | n, sentiment negation, , emoticon, TF-IDF, idiom and phrase, lexicon, adjective | |
| Khoo and Johnkhan (2018) [15] | Dataset of Amazon for reviews of products | Sentiment Lexicon | Lexicon based approach |
| John et al. (2019) [16] | Sentiment 140.com dataset | SentiWordNet (SWN) and Domain Focused (DF) lexicons. | Lexicon based approach |
| Gopi et al. (2020) [21] | Twitter dataset | TF-IDF | Machine learning based approach |
| Kanika Garg (2020) [5] | Twitter API | Hindi SentiWordNet (SWN) and AFINN-111 and Term Frequency | Lexicon based approach |
| Chauhan and Sutaria (2019) [19] | Twitter API | Semiotic lexicon | Lexicon based approach |

## IV. SOME STANDARD DATASETS USED FOR SENTIMENT ANALYSIS

Although a number of standard datasets are used for sentiment analysis, but we are discussing here commonly used dataset in this field. These datasets carry two to three classes (Positive/ Negative/ Neutral) for sentiments. Table II describes the details of standard data sets used in Sentiment Analysis.

TABLE II. DESCRIPTION OF STANDARD DATASETS FOR SENTIMENT ANALYSIS

| Dataset | Number of Tweets | Number of Classes | Description |
|---|---|---|---|
| Twitter US Airline Sentiment [17] | 14,460 | 3 | Dataset collected from Kaggle includes tweets regarding six different US (United States) airlines. It comprises 2363 positive, 9178 negative, and 3099 neutral tweets. |

| Dataset | Number of Tweets | Number of Classes | Description |
|---|---|---|---|
| IMDB Movie review dataset[18][3] | 50,000 | 2 | It is a standard dataset that includes 50k movie reviews for a binary classification task. |
| Yelp Review Dataset[4] | 10,000 | 2 | It is a yelp dataset collected from yelp used for binary classification considering 10,000 reviews with star labels. |

## V. CONCLUSION

Sentiment analysis is a technique characterized as opinion mining for categorizing text into negative class, positive class, or neutral class. It has three levels for categorizing the sentiments: aspect level, sentence level and document level. Sentiment analysis based approaches are of lexicon oriented, machine learning oriented, deep learning oriented [22] [23] and hybrid. Various applications include product analysis, movie review analysis, stock prediction, government intelligence, recommender systems, etc.

## ACKNOWLEDGMENT

This research work was funded by Institutional Fund Projects under grant no. (G:1299-611-1440). Therefore, the authors gratefully acknowledge technical and financial support from Ministry of Education and Deanship of Scientifc Research (DSR), King Abdulaziz University (KAU), Jeddah, Saudi Arabia.

## REFERENCES

[1] B. Liu, "Sentiment analysis and opinion mining," Synthesis lectures on human language technologies, vol.5, 2012, pp 1-167, https://doi.org/10.2200/S00416ED1V01Y201204HLT016

[2] P. Mehta and S. Pandya, "A review on sentiment analysis methodologies, practices and applications," International Journal of Scientific and Technology Research, vol.9, 2020, pp.601-609.

[3] IMDB Movie Review Dataset, https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews.

[4] Yelp Review Dataset, https://www.yelp.com/dataset.

[5] K. Garg, "Sentiment analysis of Indian PM's 'Mann Ki Baat'", International Journal of Information Technology, vol.12, 2020, pp.37-48.

[6] M.S. Hossen and N.R. Dev, "An Improved Lexicon Based Model for Efficient Sentiment Analysis on Movie Review Data", Wireless Personal Communication, vol.120, 2021, pp.535–544. https://doi.org/10.1007/s11277-021-08474-4

[7] H. Kumar, B. Harish, H. Darshan, "Sentiment Analysis on IMDb Movie Reviews Using Hybrid Feature Extraction Method", International Journal of Interactive Multimedia and Artificial Intelligence, vol.5,2019, pp.109–114. https://doi.org/10.9781/ijimai.2018.12.005

[8] H.K. Darshan, A.R. Shankar, B.S.Harish, H.M. Keerthi Kumar, "Exploiting RLPI for sentiment analysis on movie reviews," Journal of Advances in Information Technology, vol. 10, pp.14-19, 2019. http://www.jait.us/uploadfile/2019/0225/20190225045731879.pdf

[9] J. Kumar, J.K. Rout, A. katiyar, S.K. Jena, "Sentiment Analysis Using Weight Model Based on SentiWordNet 3.0," In: Sa P., Bakshi S., Hatzilygeroudis I., Sahoo M. (eds) Recent Findings in Intelligent Computing Techniques. Advances in Intelligent Systems and Computing, vol.709. Springer, Singapore, 2018, pp 131-139. https://doi.org/10.1007/978-981-10-8633-5_14

[10] K.S. Adewole, A.O. Balogun, M.O. Raheem, M.K. Jimoh, R.G. Jimoh, M.A. Mabayoje, F.E. Usman-Hamza, A.G. Akintola, A.W. Asaju-Gbolagade, "HYBRID FEATURE SELECTION FRAMEWORK FOR SENTIMENT ANALYSIS ON LARGE CORPORA," Jordanian Journal of Computers and Information Technology (JJCIT), vol. 7, 2021, pp.130-151.

[11] P.B. Sunitha, S. Joseph, P.V. Akhil, "A study on the performance of supervised algorithms for classification in sentiment analysis," In: TENCON-2019 IEEE Region 10 Conference (TENCON), IEEE, 2019, pp 1351-1356. https://doi.org/10.1109/TENCON.2019.8929530

[12] I. Alfina, D. Sigmawaty, F. Nurhidayati, A.N. Hidayanto, "Utilizing hashtags for sentiment analysis of tweets in the political domain", In: Proceedings of the 9th international conference on machine learning and computing, 2017, pp 43-47. https://doi.org/10.1145/3055635.3056631

[13] S. Naz, A. Sharan, N. Malik, "Sentiment classification on twitter data using support vector machine," In: 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI). IEEE, 2018, pp 676-679. https://doi.org/10.1109/WI.2018.00-13

[14] T. Eng, M.R.I. Nawab, K.M. Shahiduzzaman, "Improving accuracy of the sentence-level lexicon-based sentiment analysis using machine learning," International Journal of Scientific Research in Computer Science Engineering and Information Technology, vol.3307,2021, pp 57-68.

[15] C.S. Khoo and S.B. Johnkhan, "Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons," Journal of Information Science, vol.44, 2018, pp.491-511. https://doi.org/10.1177/0165551517703514

[16] A. John and R. Sheik,"Context deployed sentiment analysis using hybrid lexicon", In 1st International Conference on Innovations in Information and Communication Technology (ICIICT), IEEE, 2019, pp 1-5. https://doi.org/10.1109/ICIICT1.2019.8741413

[17] Airline-twitter-sentiment (2015). https://www.crowdflower.com/data/airline-twitter-sentiment

[18] A.L. Maas, R.E. Daly, P.T. Pham, D. Huang, A.Y. Ng, C. Potts, "Learning word vectors for sentiment analysis," In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol.1,2011, pp. 142–150.

[19] D. Chauhan and K. Sutaria, "Multidimensional sentiment analysis on twitter with semiotics," International Journal of Information Technology, vol.11,2019, pp.677-682.

[20] T.K. Tran and T.T. Phan, "Mining opinion targets and opinion words from online reviews," International Journal of Information Technology, vol.9,2017, pp.239-249.

[21] A.P. Gopi, R. Jyothi, V.L. Narayana and K.S. Sandeep, "Classification of tweets data based on polarity using improved RBF kernel of SVM," International Journal of Information Technology, 2020, pp.1-16.

[22] A. Aslam, A.S. Bux, Z. Habib, "Attention-based multimodal sentiment analysis and emotion recognition using deep neural networks," Applied Soft Computing, vol.144, 2023, pp.1-16.

[23] P. Atandoh, F. Zhang, D. Adu-Gyamfi, P.H. Atandoh, R.E. Nuhoho, (2023) "Integrated deep learning paradigm for document-based sentiment analysis," Journal of King Saud University–Computer and Information Sciences, vol.35,2023, pp.1-15.