An Ensemble Multi-layered Sentiment Analysis Model (EMLSA) for Classifying the Complex Datasets

Penubaka Balaji[1], D. Haritha[2]
Department of Computer Science & Engineering, K L University, Vaddeswaram, Andhra Pradesh, India

***Abstract*—Sentiment analysis is one domain that analyzes the feelings and emotions of the users based on their text messages. Sentiment analysis of short messages, reviews in online social media (OSM), and social networking sites (SNS) messages gives the analysis of given text data. Processing short text and SNS messages is a very tedious task because of the restricted detailed information generally contained. Solving this issue requires advanced techniques that are combined to give accurate results. This paper developed an Ensemble Multi-Layered Sentiment Analysis Model (EMLSA) that exploits the trust-based sentiment analysis on various real-time datasets. EMLA is the combined approach with VADER (Valence Aware Dictionary and sEntiment Reasoned) and Recurrent Neural Networks (RNNs). VADER is the lexicon and rule-based sentiment analysis model that predicts the sentiments extracted from input datasets and it is used for training. The feature extraction technique is term-frequency and inverse document frequency. Word-Level Embeddings (WLE) and Character-Level Embeddings (CLE) are the two models that increase the short text and single-word analysis. The proposed model was applied to four real-time datasets: Amazon, eBay, Trip-advisor, and IMDB Movie Reviews. The performance is analyzed using various parameters such as sensitivity, specificity, precision, accuracy, and F1-score.**

***Keywords*—Sentiment analysis; online social media; social networking sites; VADER; recurrent neural networks**

I. Introduction

Sentiment Analysis (SA) is the process of finding and dividing the opinions of the people expressed in text, voice, and videos. SA, also called opinion mining (OM), is natural language processing (NLP) that finds the emotions behind the body's text [1]. Opinions are expressed in various domains, such as movie reviews, e-commerce reviews, and Twitter reviews. Every day, many people and millions of reviews are generated by social media platforms regarding products, movies, and general topics [2]. An automated system is required to analyze the users' views, opinions, and sentiments. SA mainly focused on finding non-trivial, emotional information collected from various online sources belonging to social media [3]. Sentiment analysis can also be applied to multiple documents and phrases and analyzed single words. Finally, sentiment analysis divides the reviews into three types such as positive, negative, and neutral, based on the text data. Sentiment analysis helps e-commerce applications increase the sales of specific products [4] [5].

Natural language processing (NLP) is mainly focused on two aspects such as human language understanding and generation. It is a challenging task to analyze the natural language with the existing models. Several applications include speech recognition, text analysis, questioning and answering, synthesis of speech etc. [6]. NLP is divided into two significant areas such as sentiment analysis and recognition of emotions. Sometimes these two areas differ based on their aspects. "Emotion detection" is the domain that finds the feelings from the user's expressions like happiness, sadness, and depression. There is a significant connection between "sentiment analysis" and "emotion detection" [7]. From the emotions, the users express their feelings through text, video, and audio.

Sometimes sentiment analysis goes beyond people's opinions and views, such as sad, happy, angry, etc. [8]. Based on the feedback of the user or customer, sentiment analysis is required [9] [10] [11]. This paper describes various sentiment analyses belonging to several domains using deep learning algorithms with the integration of advanced fine-tuned models. The proposed approach focused on finding the sentiment analysis on multiple domains and analyzing the trust-based reviews in the input dataset. The proposed method also focused on aspect, Multilingual and emotion-based sentiment analysis.

II. Literature Survey

T. Gu et al. [12] proposed a novel sentiment analysis approach called MBGCV introduced to increase sentiment classification performance. MBGCV combined with various BiGRU, CNN, and VIB models. The proposed model obtained the high-level sentiment features from the given datasets. The real-time review dataset is used to analyze the performance of the proposed approach. M. K. Hayat et al. [13] introduced a DL model combined with the taxonomy-based approach to solving various issues in sentiment analytics. H. Liu [14] describes the comparative study among the lexicon, ML, and DL-based models that solve several accuracy issues. Various real-time datasets are used for experiments and analysis of sentiments. P. Gupta et al. [15] proposed the lexicon-based model that classifies the twitter data about COVID-19. The proposed model analysis the given Twitter data based on medicines, situations, and conditions faced by the users in lockdown time. This model aims to know the positive and negative opinions regarding the lockdown situation and the

performance of the Indian government. Linear SVC is used to classify the data.

A. Elouardighi et al. [16] introduced the lexicon-based model combined with N-grams and TF-IDF model. The proposed approach is applied to comments in the Arabic language collected from Facebook. The data belongs to the Legislative Elections in Morocco in 2016. Several ML algorithms are used for performance evaluation, such as NB, RF, and SVM. Effective sentiment results were analyzed by using ML algorithms. P. Vyas et al. [17] introduced the framework that works on sentiment analysis regarding COVID-19. The proposed framework extracts the positive, negative, and neutral sentiments from the Twitter data and applies various ML algorithms present for classification. R. Khan et al. [18] introduced the deep LSTM model that predicts the sentiment polarity and emotions from the sentiment140 dataset. The accuracy of proposed model is 90.23%, this is very high compare with previous models. A. S. Imran et al. [19] introduced the LSTM model for detection of emotions and sentiments in terms of text messages collected from twitter. The main drawback of this model is lack of accuracy based on several emotions such as bad, good, anger. D. Antonakaki et al. [20] describe several DL models that work on sentiment analysis. The author mainly focused on three areas such as fake news, spam content, and threats messages given on Twitter. The proposed model analyzed the better sentiments based on the result analysis—the Twitter data used for performance evaluation. H. Strobelt et al. [21] proposed a model called as LSTMVIS that process the complex patterns present in various applications. S. Kumar et al. [22] proposed a hybrid recommended system that was applied to the movies dataset. The proposed model, combined with CF and CBF, provides a better recommendation system based on sentiment analysis. The proposed approach analyzed the present trends, people's sentiments, and users' responses. S. Bhatia [23] proposed a novel graph model that analyses duplicate phrases. The proposed approach focused on correcting the sentences by using graphs. To summarize the text and reduce the dimensions, PCA is used. The proposed method achieved better opinions mining based on sentiments.

S. Davis et al. [24] discussed various works on analyzing customer reviews based on E-commerce datasets. The comparative study shows the proposed approach applied to multiple user review datasets. M. A. Tayal et al. [25] submitted an integrated system based on several operations, such as pre-processing approach. Pre-processing is used to remove ambiguity from the given dataset. The proposed method mainly combines Semantic Sentence Similarity with n-gram co-occurrence relations belonging to specific sentences. Finally, the proposed model is applied to several benchmark datasets and analyzes the performances of existing and proposed models. E. Aslanian et al. [26] proposed the hybrid recommender system (HRS) that improves the high accuracy. The proposed approach, combined with the feature relationship matrix and collaborative filtering, was used to solve the cold-start problem. The proposed method achieves better accuracy compared with other existing algorithms. E. Cambria [27] proposed an automated approach for analyzing sentiments based on emotions. The proposed system combines emotions and reviews and gives better performance.

C. Du et al. [28] proposed a new classification approach that classifies the sentiment data using an advanced feature extraction technique. The softmax classifier is adopted to increase the proposed system's performance. The F1-score of the proposed approach shows the high values for two datasets. Maria Giatsoglou et al. [29] proposed a rapid and reliable model that finds the sentiments of different types of people's opinions from other languages. The ML approach combined with the proposed approach applied to text documents initialized by vectors and trained as a polarity classification model. The proposed model is analyzed using four datasets containing reviews in Greek and English.

III. Proposed Methodology

The proposed methodology is developed with various advanced models such as the pre-trained DL model stemming model for pre-processing, TF-IDF for feature extraction, Word-Level Embeddings (WLE) for text analysis, VADER for training and RNN for classification of text data. Fig. 1 shows the step by step process of implementation.

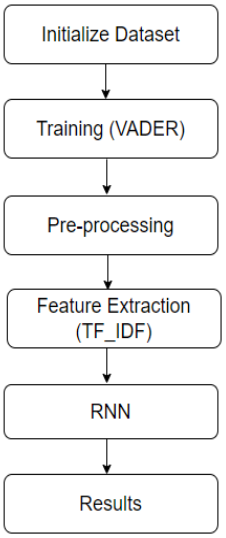

Fig. 1. System architecture.

IV. VADER

This paper uses VADER to train the given datasets to analyze the sentiments. It is the lexical database developed by using rule-based sentiment analysis. The lexicon collects the features (e.g., words) classified as positive or negative based on the sentiment polarity. VADER shows the positivity and negativity scores and also the strength of the positive and negative sentiments. The VADER is mainly based on a compound score measured by aggregation of valence scores of every word in the lexicon, find-tuned based on rules, and then normalized between -1 (high negative) and +1 (high positive). Thus this is considered the single uni-dimensional measure of sentiment for a given sentence.

$$x = \frac{x}{\sqrt{x^2 + \alpha}} \quad (1)$$

Where x = sum of valence scores of constituent words, and α = Normalization constant (default value is 15).

A. TF-IDF (Term Frequency- Inverse Document Frequency)

TF-IDF is a feature extraction approach that can extract highly reputed words in the given documents and reviews. TF mainly measures the frequently appeared mentions in the given input datasets. The term frequency refers to the total time that appeared in the given input datasets, while the document frequency refers to complete documents that contain the word. IDF counts the word from papers or reviews divided by the phrase "document frequency." Every word initialized the score by measuring the TF by its IDF. Here features mean repeated words from multiple reviews from multiple documents.

B. Word-Level Embeddings (WLE)

WLE's are encoded by using column vectors within the embedding matrix $W^{wrd} \in R^{d^{wrd} \times |V^{wrd}|}$. Every column belongs to WLE of kth word in the vocabulary. By using matrix-vector product, the word W transformed into WLE r^{wrd}.

$$r^{wrd} = W^{wrd} v^{w}) \quad (2)$$

Where v^{w} is size of vector $|V^{wrd}|$ which is value 1 at index w and 0 in all other portions. $W^{wrd} \rightarrow$ matrix. That learns and WLE size is given as d^{wrd} is hyper-parameter which is selected by the user.

C. RNN

All these layers are fully connected and are not associated with each other. RNN [30] performs better on the text sentiments dataset, and all the tasks involve sequential inputs. RNN considers one piece of information at a time and maintains the hidden units of a "state vector" consisting of data regarding the previous history based on the sequence. The outputs of hidden units are considered at various discrete time steps if the results of several neurons in a deep multi-layer network, this becomes easy to implement back-propagation to train RNN. RNN is a dynamic approach, and it is mighty to prepare them and solves the issues in back-propagated gradients either grow or shrink at each step; several times, this process is typically terminated as shown in Fig. 2.

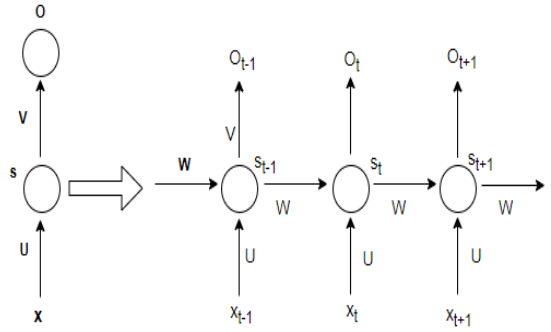

Fig. 2. Process of RNN.

The neurons in the hidden layers get the inputs from previous layers based on the time steps. Based on the above process, the RNN maps the input sequence elements represented as x_t, and the output sequence represents the elements with o_t, dependent on the previous x t′ (for $t' \leq t$). Similar metrics such as U, V, and W are utilized at every step. The back-propagation approach measures the unfolded network on the right and computes overall error based on general states s_t and all the metrics.

V. Dataset Description

1) Amazon dataset: The Amazon dataset consists of testing and training data. The training data contains 3Lakh, and testing data consists of 4Lakh data belonging to 568,000 customer's data. All these customers give reviews of the products. This is open source and free dataset collected from Kaggle: https://www.kaggle.com/datasets/bittlingmayer/amazonreviews?resource=download.

2) Ebay dataset: A data science Bootcamp project created the eBay dataset. This project aims to develop the best model for sentiment analysis. The author created this dataset using python web scraping scripts for the research work. This dataset consists of two files as ebay_reviews.csv file consists of four attributes: product category, title review, content review, and rating. The total instances are 44757. The rating attribute represents the integer value with one as the worst score and five gives the best score. The second file is a preprocessed file that consists of two attributes: rating, title review, and content review. The dataset available at: https://www.kaggle.com/datasets/wojtekbonicki/ebayreviews/discussion?select=ebay_reviews.csv.

3) Trip-advisor: This dataset consists of 20k reviews of various hotels given by customers. Trip-advisor extracts these reviews, and it is available on the Kaggle website. The dataset available at: https://www.kaggle.com/datasets/andrewmvd/tripadvisorhotel-reviews IMDB Movie Reviews Dataset: This dataset consists of 50k movie reviews and this contains 40k testing and 10 k training data. IMDB movie review dataset consists of 25k positive and 25k negative reviews and this data available at:

VI. Performance Metrics

The performance of the proposed model is analyzed by using a confusion matrix. The confusion matrix mainly measures the accurate values obtained from the proposed model. Based on the count of the importance, the performance is measured. The proposed approach focused on developing a model which can work on any dataset. The factors that show an impact on the performance of the proposed model are a true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

TP: The predicted value is true, and it is true.

TN: The predicted values is no and it is false.

FP: The predicted values is yes, originally it is not true.

FN: The predicted values is no, but the values are true.

Precision: The total positives measured from the overall positives give precision.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (3)$$

Accuracy: The model accuracy is identified.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (4)$$

Recall: The overall TPs are identified.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+ FN}} \quad (5)$$

Specificity: The overall false values are correctly identified.

$$\text{Specificity} = \frac{\text{No of TN}}{\text{No of TN+ No of FP}} \quad (6)$$

F1-Score: This measure the coherence mean of Precision and Recall achieved the better computation which is incorrectly classified cases than the Accuracy.

$$\text{F1} - \text{Score} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision+Recall})} \quad (7)$$

Result Analysis: From the results it is analyzed that the performance of various existing and proposed algorithms are given in Table I to IV. The comparative performance of RTA, IDER and EMLSA is implemented with four datasets. The proposed model EMLSA performed better on all the datasets by analyzing sentiments compared with existing models. The performance is measured by confusion matrix measures such as precision, accuracy, recall and specificity (see Fig. 3 to 6)

TABLE I. COMPARATIVE ANALYSIS OF EXISTING AND PROPOSED ALGORITHMS FOR AMAZON DATASET

Algorithms	Precision	Accuracy	Recall	Specificity	F1-Score
Reputational Trust Assessment (RTA) [31]	89.34	90.34	88.67	88.89	89.54
IDER [32]	93.21	92.23	90.45	93.89	94.9
EMLSA	98.78	98.45	97.45	98.43	98.45

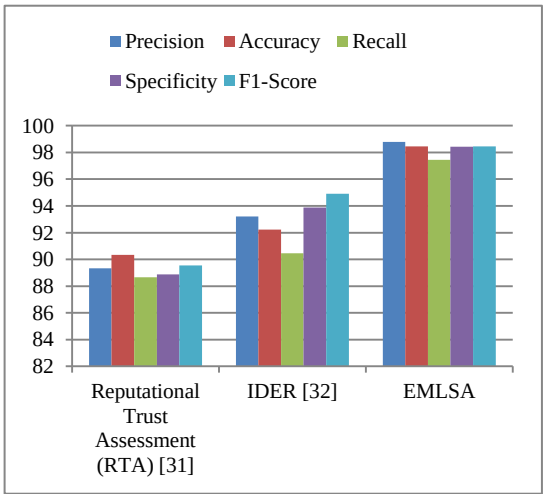

Fig. 3. Graph representation for existing and proposed algorithms for amazon dataset.

TABLE II. COMPARATIVE ANALYSIS OF EXISTING AND PROPOSED ALGORITHMS FOR EBAY DATASET

Algorithms	Precision	Accuracy	Recall	Specificity	F1-Score
Reputational Trust Assessment (RTA) [31]	87.12	88.56	87.23	87.34	88.65
IDER [32]	92.34	92.56	92.98	92.67	93.5
EMLSA	97.34	97.78	98.34	97.65	97.12

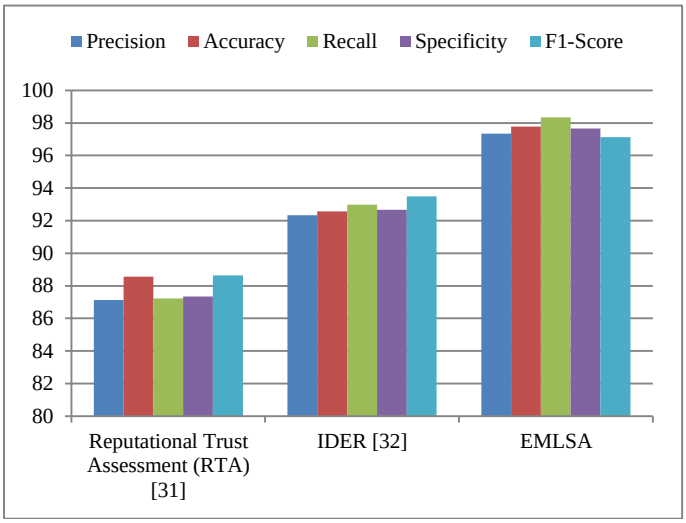

Fig. 4. Graph representation for existing and proposed algorithms for eBay dataset.

TABLE III. COMPARATIVE ANALYSIS OF EXISTING AND PROPOSED ALGORITHMS FOR TRIP-ADVISOR

Algorithms	Precision	Accuracy	Recall	Specificity	F1-Score
Reputational Trust Assessment (RTA) [31]	88.96	89.12	88.32	88.67	88.32
IDER [32]	93.45	94.67	93.87	93.45	93.76
EMLSA	99.23	99.56	99.22	98.11	98.78

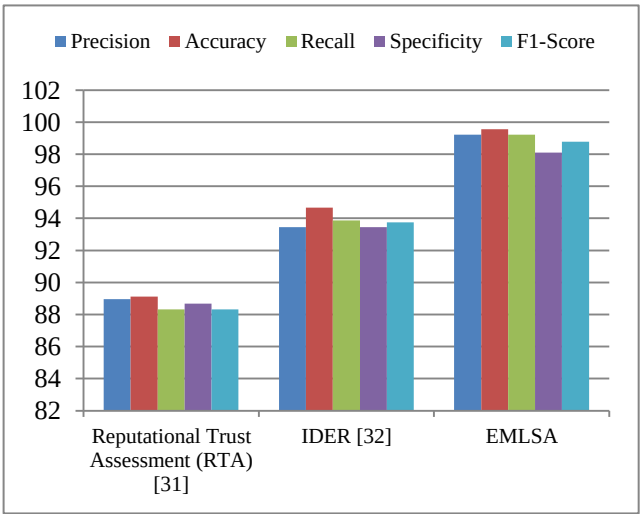

Fig. 5. Graph representation for existing and proposed algorithms for trip-advisor.

TABLE IV. PERFORMANCE OF EXISTING AND PROPOSED ALGORITHMS FOR IMDB MOVIE REVIEWS DATASET

Algorithms	Precision	Accuracy	Recall	Specificity	F1-Score
Reputational Trust Assessment (RTA) [31]	91.89	91.78	91.45	91.99	91.23
IDER [32]	94.34	94.1	94.9	94.6	94.67
EMLSA	99.8	99.8	99.7	99.3	99.43

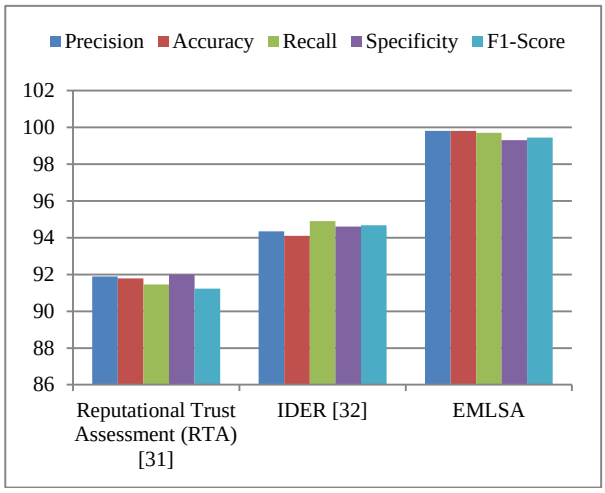

Fig. 6. Graph representation for existing and proposed algorithms for IMDB movie reviews dataset.

VII. CONCLUSION

Even though a conclusion may review the main results this paper describes the new DL model that can process complex datasets based on the reviews given by the users. The proposed approach was applied to four benchmark datasets that show the comparative performance in terms of sensitivity, specificity, accuracy, precision, and f1-score. The proposed DL model focused on extracting every aspect of the input reviews. The word embedding models TF-IDF, and Word2Vec combined with the DL model give high performance in terms of given input datasets. The proposed model achieved an accuracy of 98.45% for the amazon dataset, 97.78% for the TripAdvisor dataset, and 99.56% for the ebay dataset and for IMDB dataset the accuracy is 99.8%. Thus it is shown that the accuracy is more for the proposed model. In future, the multi-layered models are to be developed by improving the sentiments and emotion detection. Various combined and integrated models are required to increase the performance.

REFERENCES

[1] Cambria E, Dragoni M, Kessler B, Donadello I. Ontosenticnet 2: enhancing reasoning within sentiment analysis. IEEE Intell Syst. 2022;37(2):103–110.

[2] Cambria E, Xing F, Thelwall M, Welsch R. Sentiment analysis as a multidisciplinary research area. IEEE Trans Artif Intell. 2022;3(2):1–4.

[3] Chan JYL, Bea KT, Leow SMH, Phoong SW, Cheng WK. State of the art: a review of sentiment analysis based on sequential transfer learning. Artif Intell Rev. 2022 doi: 10.1007/s10462-022-10183-8.

[4] Cheng WK, Bea KT, Leow SMH, Chan JY-L, Hong Z-W, Chen Y-L. A review of sentiment, semantic and event-extraction-based approaches in stock forecasting. Mathematics. 2022;10(14):2437. doi: 10.3390/math10142437.

[5] Da'u A, Salim N, Rabiu I, Osman A. Recommendation System Exploiting Aspect-Based Opinion Mining with Deep Learning Method. Inf Sci. 2020;512:1279–1292. doi: 10.1016/j.ins.2019.10.038.

[6] Itani M, Roast C, Al-Khayatt S (2017) Developing resources for sentiment analysis of informal Arabic text in social media. Procedia Comput Sci 117:129–136.

[7] Munezero M, Montero CS, Sutinen E, Pajunen J (2014) Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. IEEE Trans Affect Comput 5(2):101–111.

[8] Penubaka Balaji, D Haritha, "Feature Based Summarization System for E-Commerce Based Products by Using Customer's Reviews," CCODE-2018, http://dx.doi.org/10.2139/ssrn.3168342.

[9] Penubaka Balaji, O. Nagaraju, D Haritha, "Levels of sentiment analysis and Its challenges", ICBDACI-2017,doi:10.1109/icbdaci.2017. 8070879

[10] Penubaka Balaji, O. Nagaraju, D Haritha, "CommuTrust: Reputation based trust evaluation in E-Commerce applications", ICBDACI-2017, doi: 10.1109/icbdaci.2017.8070856.

[11] Penubaka Balaji, O. Nagaraju, D Haritha, "An Overview on Opinion Mining Techniques and Sentiment Analysis", IJPAM, Vol.118 & Issu.19, Jan 2018.

[12] M.K. Rahmat, S. Jovanovic, K.L. Lo, Reliability and Availability T. Gu, G. Xu and J. Luo, "Sentiment Analysis via Deep Multichannel Neural Networks With Variational Information Bottleneck," in IEEE Access, vol. 8, pp. 121014-121021, 2020, doi: 10.1109/ACCESS.2020.3006569.

[13] M. K. Hayat et al., "Towards Deep Learning Prospects: Insights for Social Media Analytics," in IEEE Access, vol. 7, pp. 36958-36979, 2019, doi: 10.1109/ACCESS.2019.2905101.

[14] H. Liu, I. Chatterjee, M. Zhou, X. S. Lu and A. Abusorrah, "Aspect-Based Sentiment Analysis: A Survey of Deep Learning Methods," in IEEE Transactions on Computational Social Systems, vol. 7, no. 6, pp. 1358-1375, Dec. 2020, doi: 10.1109/TCSS.2020.3033302.

[15] P. Gupta, S. Kumar, R. R. Suman and V. Kumar, "Sentiment Analysis of Lockdown in India During COVID-19: A Case Study on Twitter," in IEEE Transactions on Computational Social Systems, vol. 8, no. 4, pp. 992-1002, Aug. 2021, doi: 10.1109/TCSS.2020.3042446.

[16] A. Elouardighi, M. Maghfour, H. Hammia and F. -z. Aazi, "A machine Learning approach for sentiment analysis in the standard or dialectal Arabic Facebook comments," 2017 3rd International Conference of Cloud Computing Technologies and Applications (CloudTech), 2017, pp. 1-8, doi: 10.1109/CloudTech.2017.8284706.

[17] P. Vyas, M. Reisslein, B. P. Rimal, G. Vyas, G. P. Basyal and P. Muzumdar, "Automated Classification of Societal Sentiments on Twitter With Machine Learning," in IEEE Transactions on Technology and Society, vol. 3, no. 2, pp. 100-110, June 2022, doi: 10.1109/TTS.2021.3108963.

[18] R. Khan, R. Khan, P. Shrivastava, A. Kapoor, A. Tiwari and A. Mittal, "Social media analysis with AI: Sentiment analysis techniques for the analysis of Twitter COVID-19 data", J. Crit. Rev., vol. 7, no. 9, pp. 2761-2774, 2020.

[19] A. S. Imran, S. M. Doudpota, Z. Kastrati and R. Bhatra, "Cross-cultural polarity and emotion detection using sentiment analysis and deep learning—A case study on COVID-19", IEEE Access, vol. 8, pp. 181074-181090, 2020.

[20] D. Antonakaki, P. Fragopoulou and S. Ioannidis, "A survey of Twitter research: Data model graph structure sentiment analysis and attacks", Expert Syst. Appl., vol. 164, Feb. 2021, [online] Available: https://doi.org/10.1016/j.eswa.2020.114006.

[21] H. Strobelt, S. Gehrmann, H. Pfister, And A. M. Rush, "Lstmvis: A Tool For Visual Analysis Of Hidden State Dynamics In Recurrent Neural Networks," Ieee Trans. Vis. Comput. Graph., 2018.

[22] S. Kumar, K. De and P. P. Roy, "Movie Recommendation System Using Sentiment Analysis From Microblogging Data," in IEEE Transactions on Computational Social Systems, vol. 7, no. 4, pp. 915-923, Aug. 2020, doi: 10.1109/TCSS.2020.2993585.

[23] S. Bhatia, "A Comparative Study of Opinion Summarization Techniques," in IEEE Transactions on Computational Social Systems, vol. 8, no. 1, pp. 110-117, Feb. 2021, doi: 10.1109/TCSS.2020.3033810.

[24] S. Davis and N. Tabrizi, "Customer Review Analysis: A Systematic Review," 2021 IEEE/ACIS 6th International Conference on Big Data, Cloud Computing, and Data Science (BCD), 2021, pp. 91-97, doi: 10.1109/BCD51206.2021.9581965.

[25] M. A. Tayal, M. M. Raghuwanshi and L. G. Malik, "ATSSC: Development of an approach based on soft computing for text summarization" in Comput. Speech Lang., vol. 41, pp. 214-235, Jan. 2017.

[26] E. Aslanian, M. Radmanesh and M. Jalili, "Hybrid recommender systems based on content feature relationship", IEEE Trans. Ind. Informat., Nov. 2016.

[27] E. Cambria, "Affective computing and sentiment analysis", IEEE Intell. Syst., vol. 31, no. 2, pp. 102-107, Mar./Apr. 2016.

[28] C. Du and L. Huang, "Text classification research with attention-based recurrent neural networks", Int. J. Comput. Commun. Control, vol. 13, no. 1, pp. 50-61, 2018.

[29] Maria Giatsoglou, Manolis G. Vozalis, Konstantinos Diamantaras, Athena Vakali, George Sarigiannidis, Konstantinos Ch. Chatzisavvas, Sentiment analysis leveraging emotions and word embeddings, Expert Systems with Applications, Volume 69, 2017, Pages 214-224, https://doi.org/10.1016/j.eswa.2016.10.043.

[30] Wang H, Raj B, Xing E P. On the origin of deep learning. 2017.

[31] Penubaka, Balaji & Haritha, D.. (2018). Opinion Mining Based Reputational Trust Assessment in E-Commerce Applications. Journal of Advanced Research in Dynamical and Control Systems. 10.

[32] Donavalli, H. & Penubaka, Balaji. (2019). Identification of opinionated features extraction from unstructured textual reviews. International Journal of Recent Technology and Engineering. 7. 674-677.