Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5769 - 5778

Received 20 May 2021; Accepted 25 May 2021.

Predictive Analytics for Sentiment Classification of Social Media Data Using Deep Neural Network

Savitha Hiremath[1], S H Manjula[2] and K R Venugopal[3]

[1]Assistant Professor, Department of ISE, CMR Institute of Technology, Bengaluru-560037

[2] Professor, Department of CSE, UVCE, Bangalore University, Bengaluru-560001

[3]Vice-Chancellor, Bangalore University, Bengaluru-560056

[1]savitha.h@cmrit.ac.in

Abstract

A huge amount of user-generated data in the form of tweets or reviews on social media can be collected and analyzed for making informed decisions. This paper uses the novel deep learning model, namely the Elite Opposition-based Bat Algorithm for Deep Neural Network (EOBA-DNN) for performing polarity classification of the social media data. The proposed method includes three major steps, such as preprocessing, term weighting, and sentiment classification for identifying the polarity of the data. The results show that the EOBA-DNN outperforms other existing algorithms with improved accuracy for Sentiment Classification.

Keywords:
Bat Algorithm; Sentiment Classification; Deep Neural Network (DNN); Predictive analytics; Sentiment analysis; Social media; Twitter data; Tem weighting.

1. INTRODUCTION

The way populace consume, converse, collaborate together with create is being revolutionized by Social Media (SM) [1, 2, 3]. Technologies categorized as "social media" contain online sites namely Facebook, Instagram, along with Twitter [4], message boards, blogs, etc [5]. For every firm, several benefits are brought by social media, namely the capability of extending to bigger audiences, the competence of driving sales via social commerce, accompanied by the skill to create trust and reputation [6]. A rising method for comprehending the individual’s opinions through social networks is Sentiment Analysis. [7].

When implemented to the comments of Social Media network users, their views are ascribed as being positive, neutral, or negative [8]. SA is applied by numerous preceding studies [9, 10] into a goods or movie review for deeper understanding their customer and making the required decision to ameliorate their product or services [11]. For conducting Sentiment Analysis, Lexicon along with Machines Learning (ML)-centered approaches is utilized. When informal language is employed, Lexicon-centered algorithms lose their efficacy [12]. Generally, huge datasets are processed utilizing ML classifiers. But, physical labeling of text aimed at ML is difficult and tedious to perform [13]. Deep Neural Networks (DNN) assists the algorithms to study by resolving the issue of training the complex models with comparatively bigger datasets with the progression in Artificial Intelligences [14]. A DL is proposed here, namely EOBA-DNN for executing Sentiment Classifications (SC) of Social media data.

Received 20 May 2021; Accepted 25 May 2021.

This paper has been categorized as: Section 2 provided the associated work. Section 3 briefly describes the proposed work. Section 4 exhibits the proposed method's results together with a discussion. Lastly, a conclusion is delineated in section 5.

2. LITERATURE REVIEW

Nadia Chouchani *et al.*, [15] introduced a polarity classification method for executing SA of SM data by exploring social network structures. The information regarding social influence processes had improved SA. The model had surpassed an approach that deemed only information concerning homophily as revealed by the results. But, only on a smaller twitter dataset, the technique was tested.

Sanur Sharma and Anurag Jain [16] introduced a Twitter SA by crawling real-time Twitter data and implementing ML techniques for effectively classifying the data. The tweet sentiments were effectively classified by the results into positive as well as negative with 87.2% accuracy. But, immense datasets were needed by the ML algorithms for system training and were extremely liable to errors.

Srishti Vashishtha and Seba Susan [17] provided a SA of SM posts utilizing a compilation of fuzzy rules including multiple lexicons and datasets. Higher performance was yielded by the experiments upon benchmark datasets for the approach as analogized to the top-notch. But, the fuzzy logic was not always precise. Therefore, grounded upon the assumption, the outcomes were perceived so it mightn't be extensively accepted.

SoYeop Yoo *et al.*, [18] offered a system aimed at examining and predicting users' sentimental trajectories for events analyzed in real-time out of the enormous SM contents. The SA's accuracy was ameliorated by employing the DL technique as demonstrated by the results. But, the training time was improved by the DL algorithms without optimization and the traditional use of loss function caused miss classification error.

3. PROPOSED METHODOLOGY

A Deep Learning approach, such as EOBA-DNN, is proposed aimed at executing Sentiment Analysis of data. As of the publicly accessible database, the SM data is gathered initially. For eliminating the noise on the inputted data, preprocessing operations are executed. Next, utilizing a weighting scheme, like TFIDF-DFS, Feature Extractions (FE) of the preprocessed data is done. Lastly, for SC, the features (extracted) of the inputted data are rendered to the EOBA-DNN that classified the sentiment of provided inputted data into positive, negative, along with neutral. The proposed method's framework is exhibited in Figure 1.

Received 20 May 2021; Accepted 25 May 2021.

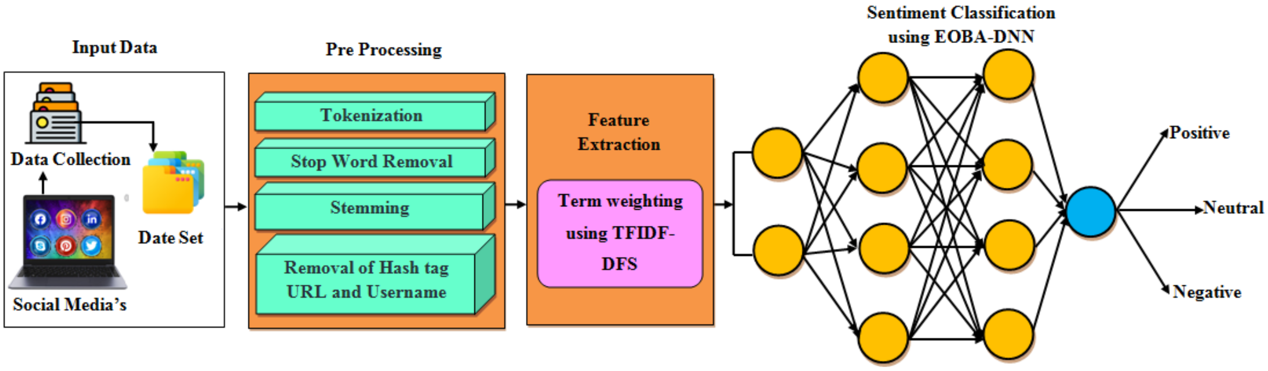


Figure 1: Proposed framework

3.1 Preprocessing

An essential procedure for SA which includes cleaning and filtering of data is Pre-processing. Lots of noisy information, say, mis-spelled, slang words, user-generated abbreviations, and also white spaces that could demean the classifier's accuracy is contained by the amassed data of an openly available database. Therefore, preprocessing is performed, which encompasses (a) tokenization: a whole sentence is separated into smaller units named tokens, (b) stop word removal: the frequently utilized words are removed which are not likely to be obliging for learning, (c) stemming: a word is lessened to its word stem that affixes to suffixes and prefixes and (d) Removal of Hashtag, URL, and Username: the hashtags marked using a number sign (#) before a word on a sentence are removed and if an URL or username exists on a sentence, it is eliminated.

3.2 Feature extraction

FE is performed after preprocessing. A weight is allotted for every term utilizing the TFIDF-DFS scheme in the FE stage. Distinguishing feature selector (DFS) allots scores to every term deeming their distinguishing power. Terms frequency (TF) is a raw term frequency (total times a term takes place within a document), together with Inverse Documents Frequency (IDF) is the logarithmically scaled inverse part of the documents that encompass the word (acquired by division of the complete documents with the total documents encompassing the term with the logarithm of that quotient is taken). The paper utilizes the unification of these '3' techniques, which is expressed in Equation (1).

$$W_{TFIDF-DFS} = TF\left(t_i, d_k\right) \times \left(1 + \log\left(\frac{D}{d\left(t_i\right)}\right)\right) \times \sum_{z=1}^{M} \left(\frac{\left(\frac{t_i c_z}{t_i c_z + t_i \bar{c}_z}\right)}{\left(\frac{\bar{t}_i c_z}{t_i c_z + \bar{t}_i c_z}\right) + \left(\frac{t_i \bar{c}_z}{t_i \bar{c}_z + \bar{t}_i \bar{c}_z}\right) + 1} \right) \tag{1}$$

Wherein M indicates the total classes on the amassed data, $TF\left(t_i, d_k\right)$ signifies the occurrence frequency of term t_i in the document d_k, the 2nd term in equation (1) signifies the

IDF (class frequency) of term t_i, and 3rd term symbolizes the DFS scheme. In the 3rd term, $t_i c_z$ denotes the class c_z including the term t_i, $\bar{t}_i c_z$ is the class c_z does not contain the term t_i, $t_i \bar{c}_z$ signifies that the term t_i is not a member of the class c_z, and $\bar{t}_i \bar{c}_z$ implies the total sentences that did not comprise the term t_i in other classes.

3.3 Classification

The EOBA-DNN is engaged for polarity classification after TW. The DNN [19][20] comprises an input layer, an output layer (OL), and above '1' Hidden Layer (HL) in between. The DNN's weight learning is deemed as a heavy complicated optimizing process of parameter system. EOBA is applied for increasing the computation speed and prediction's accuracy of DNN that encompasses a great ability for global searching, and rapidly converging is employed efficiently to ascertain the weights corresponding to different connections. Therefore, the proposed work is termed EOBA-DNN. The algorithmic procedure of DNN is provided as follows

Step 1: Get the term weight of every term acquired by TFIDF-DFS on the dataset, which is taken as the input.

Step 2: Create optimized weights values for each data (input) that is rendered on the input layer using EOBA and allocate it to the HL neurons in tandem with the OL neurons.

Step 3: Gauge the HL's output using equations (2) and (3). The proposed work utilizes '2' HLs for polarity classification.

$$H_{1i} = b_{1i} + \sum_{i=1}^{n} T_i w_{1i} \tag{2}$$

$$H_{2i} = b_{2i} + \sum_{i=1}^{n} H_{1i} w_{2i} \tag{3}$$

Wherein $b_{1i}, and\ b_{2i,}$ and signifies the 1st and 2nd HL's bias values, $w_{1i}$and $\mathrm{w}_{2\mathrm{i}}$ indicates the optimized weight values of the HL '1' and '2', and T_i signifies the input data values as of the earlier step.

Step 4: Assess the output by multiplication of the final HL with the weight of the same HL's output that is exhibited in equation (4).

$$P_J = G\left(b_i + \sum_{i=1}^{n} H_{2i} w_i \right) \tag{4}$$

Wherein, $G(\cdot)$ signifies the nonlinear activation function, w_i signifies the optimized weight of the final HL's output, and b_i represents the bias value of the last HL's output and P_i signifies the output unit. The EOBA-DNN's output indicates '3' data classes: positive, negative, along with neutral.

Received 20 May 2021; Accepted 25 May 2021.

Step 5: Calculate the OL's activation function with the learning error is envisaged by the loss function. The activation along with loss function engaged on the proposed EOBA-DNN are ReLU (ensured no negative outputs and ReLU surpassed the other activity functions owing to less training- and validating-time) along with cross-entropy which are formulated in equations (5) and also (6)

$$G(T_w) = \max(0, T) \tag{5}$$

$$C_{El} = -\sum_{i=1}^{n} \sum_{j=1}^{J} \left(F_{j,i} - P_{j,i} \right) \tag{6}$$

Wherein, n signifies the number of data points (inputs), J signifies the total outputs, $F_{j,i}$ implies the j th target value for data point i, and $P_{j,i}$ indicates the j th output attained for data point i. The weight optimization executed by DNN employing EOBA is elucidated in the below section.

3.3.1 EOBA for weight optimization of DNN

Centered upon the microbats' echolocation behavior, the Bat Algorithms (BA) [21] was inspired through differing pulse rates of emission along with loudness. The algorithm includes the drawbacks of falling into a local optimum solution along with premature convergence. For overcoming this, an elite opposition-centered (EO) learning approach is added in the BA. An efficient search mechanism that could augment population diversity and improve global searchability is the EO. Therefore, the algorithm is labeled EOBA.

In a D-dimensional Search Space (SS), once the time is t, the frequency (f_i), the position is $(T)_i^t$, the bat's speed (v_i^t) in the populace are updated as exhibited in equations (7) - (9).

$$f_i = f_{\min} + (f_{\max} - f_{\min})\lambda \tag{7}$$

$$v_i^t = v_i^{t-1} + \left((T)_i^{t-1} - (T)_i^t \right) f_i \tag{8}$$

$$(T)_i^t = (T)_i^{t-1} + v_i^t \tag{9}$$

Wherein, $\lambda \in (0,1)$ is an arbitrary variable that complies with uniform distribution; here T denotes the existing global optimum position. In the local search, the updated local position formula is provided as follows:

$$(T)_{new} = (T)_{old} + \chi(Q^t) \tag{10}$$

Where $(T)_i^t$ symbolizes the existing optimal solution set, Q^t signifies the mean response of bats at the same time period and χ denotes a random value. Pulse loudness (Q^t) along with pulse rate S_i is constantly updated throughout the iteration. The updated formula is:

$$Q_i^{t+1} = \alpha \times Q_i^t \tag{11}$$

$$S_i^{t+1} = S_i^0 (1 - \exp(\beta t)) \tag{12}$$

Where α and β are constants and $0 < \alpha < 1, \beta > 0$. When $t \to \infty$, $S_i^t \to 0, and\ S_i^t = S_i^0$. After contrasting the fitness value of the practicable solution with the inverse solution of each bat, the superior individual is deliberated as elite bat $T_e = T_{e1}, T_{e2} ....T_{e.D}$. The bat $(T)_i^t$ and elite inverse solution $(\hat{T})_i^t$ are $(T)_i^t = (T)_{i,1}^t, (T)_{i,2}^t, ...(T)_{i,D}^t$ and $(\hat{T})_i^t = (\hat{T})_{i,1}^t, (\hat{T})_{i,2}^t, ...(\hat{T})_{i,D}^t$ correspondingly, and the formula is exhibited in equation (13).

$$(\hat{T})_i^t = k.(xa_j + xb_j) - T_{e,j},\ where\, i = 1,2,...n,\ j = 1,2,....D \tag{13}$$

Wherein n signifies the population's size, D is the SS dimension, k ∈ (0, 1), and xa_j and xb_j imply the dynamic boundaries of j th decision variable, which are gauged as exhibited in equations (14) and (15).

$$xa_j = \min(T_{i,j}^t) \tag{14}$$

$$xb_j = \max(T_{i,j}^t) \tag{15}$$

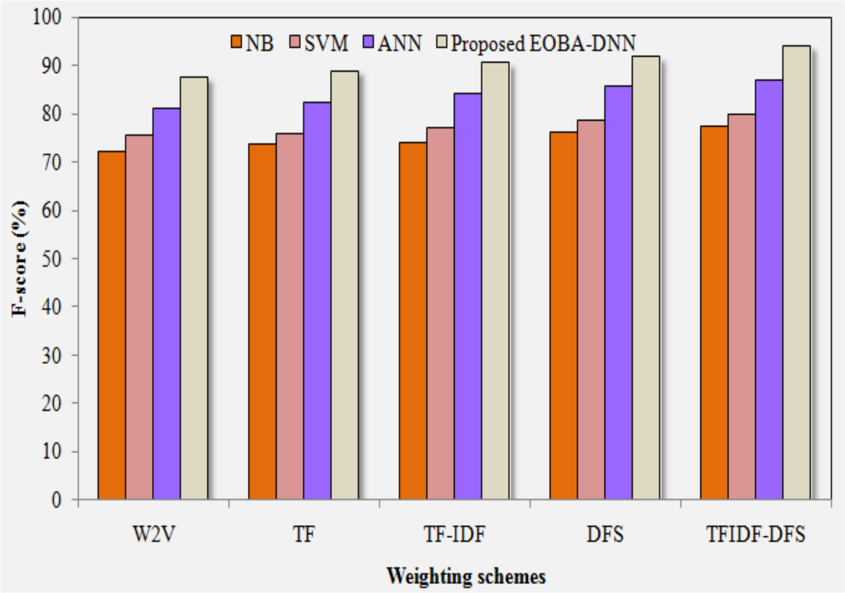
The fixed boundary is changed by the dynamic boundary of the SS, which is employed for preserving the optimum solution. The inverse solution jumps out xa_j, xb_j is deemed as a feasible solution in equation (16).

$$(\hat{T})_{i,j}^t = rd(xa_j, xb_j),\ if\ (\hat{T})_{i,j}^t < xa_j\ or\ (\hat{T})_{i,j}^t > xb_j \tag{16}$$

4. RESULTS AND DISCUSSION

The proposed EOBA-DNN for SC of SM data is executed in Python and the publicly accessible Twitter dataset is employed here for examining the results. The proposed EOBA-DNN's results with the existent classifiers namely Support Vectors Machine, Naive Bayes (NB), along with Artificial Neural Networks concerning f-measure and accuracy was examined by this section [22]. The technique's f-measure together with accuracy are contrasted with different weighting schemes, namely Word to Vector (W2V), TF, TF-IDF, DFS, and proposed TFIDF-DFS, which is displayed in Figure 2.

Received 20 May 2021; Accepted 25 May 2021.



(a)

Figure 2 (a): Results of proposed and existing classifiers

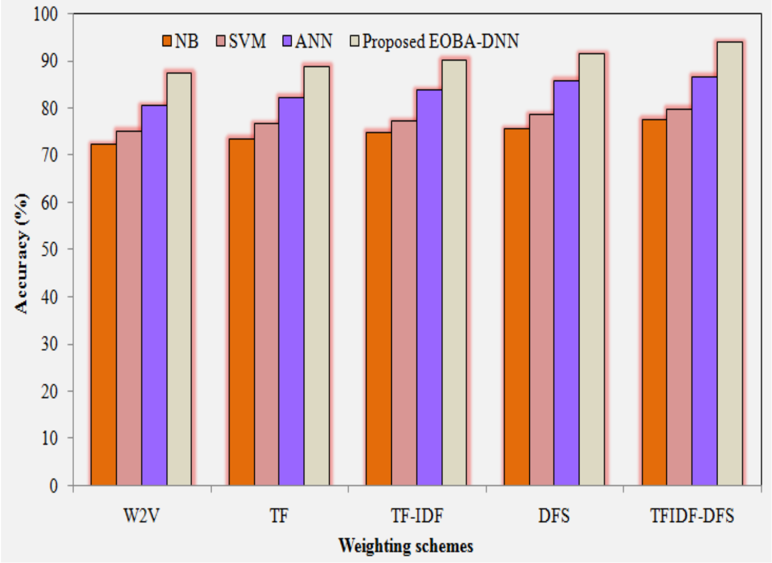


Figure 2 (b): Results of proposed and existing classifiers

Figure 2 (a) and (b) confirms that the proposed EOBA-DNN attains the highest level of accuracy and f-score when compared to other algorithms. It attains the accuracy values of 87.56, 88.95, 90.25, 91.58, and 93.99 and attains the f-score values of 87.46, 88.85, 90.45, 91.68, and 93.99 for the weighting schemes W2V, TF, TFIDF, DFS, and TFIDF-DFS. The NB attains the very lowest value of f-score and accuracy for all weighting schemes. The ANN attains an average level of performance when compared to all. Comparing all weighting schemes, W2V

Received 20 May 2021; Accepted 25 May 2021.

attains the very lowest performance for all classifiers and the proposed weighting scheme attains the highest level of performance for all classifiers. So the results of techniques finally show that when using the proposed weight scheme, the polarity classification of the Twitter data is performed more accurately.

5. CONCLUSION

This paper proposes an Elite Opposition-based Bat Algorithm for Deep Neural Network (EOBA-DNN) classifier for Sentiment Analysis of social media data that can predict the future events of the particular domain. With the polarity classification (positive, negative, and neutral) of the social media data, the decision-making process has been done in various applications, such as business, politics, government intelligence, summarization, recommender system, etc. The current EOBA-DNN achieved improved accuracy when compared to other classifiers and when using the proposed term weighting approach the classification techniques attain the highest level of the f-score and accuracy. In the future, this work has been extended to develop a deep learning model with the feature selection algorithm for performing sentiment classification of both textual and visual data.

REFERENCES

[1] Jaewoong Choi, Janghyeok Yoon, Jaemin Chung, Byoung-Youl Coh and Jae-Min Lee, “Social media analytics and business intelligence research a systematic review”, Information Processing & Management, vol. 57, no. 6, pp. 102279, 2020.

[2] Byeongki Jeong, Janghyeok Yoon and Jae-Min Lee, “Social media mining for product planning a product opportunity mining approach based on topic modeling and sentiment analysis”, International Journal of Information Management, vol. 48, pp. 280-290, 2019, 10.1016/j.ijinfomgt.2017.09.009.

[3] Yuanzhu Zhan, Runyue Han, Mike Tse, Mohd Helmi Ali and Jiayao Hu, “A social media analytic framework for improving operations and service management a study of the retail pharmacy industry”, Technological Forecasting and Social Change, vol. 163, pp. 120504, 2021, 10.1016/j.techfore.2020.120504.

[4] Samah Mansour, “Social media analysis of user’s responses to terrorism using sentiment analysis and text mining”, Procedia Computer Science, vol. 140, pp. 95-103, 2018, 10.1016/j.procs.2018.10.297.

[5] Lindsay E Young, Stephanie Soliz, Jackie Jingyi Xu and Sean Young, “A review of social media analytic tools and their applications to evaluate activity and engagement in online sexual health interventions”, Preventive Medicine Reports, pp. 101158, 2020, 10.1016/j.pmedr.2020.101158.

[6] Sang Hoon Jung and Yong Jin Jeong, “Twitter data analytical methodology development for prediction of start-up firms social media marketing level”, Technology in Society, vol. 63, pp. 101409, 2020, 10.1016/j.techsoc.2020.101409.

Received 20 May 2021; Accepted 25 May 2021.

[7] Muhammad Alam, Fazeel Abid, Cong Guangpei and Yunrong L. V, “Social media sentiment analysis through parallel dilated convolutional neural network for smart city applications”, Computer Communications, vol. 154, pp. 129-137, 2020, 10.1016/j.comcom.2020.02.044.

[8] Karen Howells and Ahmet Ertugan, “Applying fuzzy logic for sentiment analysis of social media network data in marketing”, Procedia Computer Science, vol. 120, pp. 664-670, 2017, 10.1016/j.procs.2017.11.293.

[9] Filippo Chiarello, Andrea Bonaccorsi and Gualtiero Fantoni, “Technical sentiment analysis Measuring advantages and drawbacks of new products using social media”, Computers in Industry, vol. 123, pp. 103299, 2020, 10.1016/j.compind.2020.103299.

[10] Ali Derakhshan and Hamid Beigy, “Sentiment analysis on stock social media for stock price movement prediction”, Engineering Applications of Artificial Intelligence, vol. 85, pp. 569-578, 2019, 10.1016/j.engappai.2019.07.002.

[11] Zulfadzli Drus and Haliyana Khalid, “Sentiment analysis in social media and its application systematic literature review”, Procedia Computer Science, vol. 161, pp. 707-714, 2019, 10.1016/j.procs.2019.11.174.

[12] Pietro Ducange, Michela Fazzolari, Marinella Petrocchi and Massimo Vecchio, “An effective Decision Support System for social media listening based on cross-source sentiment analysis models”, Engineering Applications of Artificial Intelligence, vol. 78, pp. 71-85, 2019, 10.1016/j.engappai.2018.10.014.

[13] Muhammad Asif, Atiab Ishtiaq, Haseeb Ahmad, Hanan Aljuaid and Jalal Shah, “Sentiment analysis of extremism in social media from textual information”, Telematics and Informatics, vol. 48, pp. 101345, 2020, 10.1016/j.tele.2020.101345.

[14] Fazeel Abid, Chen Li and Muhammad Alam, “Multi-source social media data sentiment analysis using bidirectional recurrent convolutional neural networks”, Computer Communications, vol. 157, pp. 102-115, 2020, 10.1016/j.comcom.2020.04.002.

[15] Nadia Chouchani and Mourad Abed, “Enhance sentiment analysis on social networks with social influence analytics”, Journal of Ambient Intelligence and Humanized Computing, vol. 11, no. 1, pp. 139-149, 2020.

[16] Sanur Sharma and Anurag Jain, “Cyber social media analytics and issues a pragmatic approach for twitter sentiment analysis”, In Advances in Computer Communication and Computational Sciences, Springer, Singapore, pp. 473-484, 2019, 10.1007/978981-13-6861-5_41.

[17] Srishti Vashishtha and Seba Susan, “Fuzzy rule based unsupervised sentiment analysis from social media posts”, Expert Systems with Applications, vol. 138, pp. 112834, 2019, 10.1016/j.eswa.2019.112834.

Received 20 May 2021; Accepted 25 May 2021.

[18] SoYeop Yoo, JeIn Song and OkRan Jeong, "Social media contents based sentiment analysis and prediction system", Expert Systems with Applications, vol. 105, pp. 102-111, 2018, 10.1016/j.eswa.2018.03.055.

[19] Savitha Mathapati, Ayesha Nafeesa, Tanuja R, S H Manjula and Venugopal K R, "Semi-supervised Domain Adaptation and Collaborative Deep Learning for Dual Sentiment Analysis", SN Applied Sciences, Springer. July 2019 .

[20] JayaLakshmi A. N. M and Krishna Kishore K V, "Performance evaluation of DNN with other machine learning techniques in a cluster using Apache Spark and MLlib", Journal of King Saud University-Computer and Information Sciences, 10.1016/j.jksuci.2018.09.022, 2018.

[21] Hongwei Tang, Wei Sun, Hongshan Yu, Anping Lin and Min Xue, "A multirobot target searching method based on bat algorithm in unknown environments", Expert Systems with Applications, vol. 141, pp. 112945, 2020, 10.1016/j.eswa.2019.112945.

[22] Savitha Mathapati, Ayesha Nafeesa, S H Manjula and Venugopal K R, "Semi-supervised Cross Domain Sentiment Classification on Tweets using Optimized Topic-Adaptive Word Expansion Technique", International Journal of Computer Science and Information Security (IJCSIS), vol. 15, no. 5, pp. 370-381, May 2017.