Section A-Research paper

Machine Learning Algorithms in Big Data Analytics for social media data based sentimental analysis

Yogendra Narayan Prajapati[*1] , Dr.U.Sesadri[2], Mahesh T R[3], Shreyanth S[4], Dr.Ashish Oberoi[5] Dr Khel Prakash Jayant [6]

Abstract: Due to the extensive usage of the Internet, social media has grown to play a significant role in our daily lives. Twitter is one of the most popular social media platforms in use today. People express their ideas through tweets on a variety of topics, including politics, sports, the economy, and more. The massive dataset produced by the daily millions of tweets caught the interest of data scientists, who decided to concentrate on it for sentiment analysis. This research propose novel technique in machine learning algorithm in big data for social media based sentimental analysis. Here the input data has been collected as social media based sentimental data and processed for noise removal. Then this data has been clustered using Fuzzy-C means clustering and feature extracted using differential multi-layer whale optimization. The experimental analysis has been carried out in terms of accuracy, precision, recall, AUC and RMSE. The proposed technique attained accuracy of 95%, precision of 72%, recall of 62%, AUC of 44%, RMSE of 52%.

Keywords: *Social media, machine learning algorithm, big data, sentimental analysis, optimization, clustering*.

1. Introduction

With the significant development of social media, the web appears to be a vibrant and lively space where billions of people from all over the world engage, share, post, and carry out a variety of daily activities.People can connect and communicate with one another through social media at any time and from anywhere [1].Social media offers numerous ways for individuals to express their opinions on recent or historical events as well as numerous other activities going on in our world. On a daily basis, the web is visited by more than 500 million people worldwide [2].Many different social media platforms produce a significant amount of knowledge in many different formats and languages. Finding the new purpose and obtaining information from it provide hurdles for data analytics.This enormous volume of data can be converted into information to aid in operational, managerial, and strategic decision-making by using the proper methods and techniques.The analysis of ordinary material, such as that found in blogs, open forums, and traditional review channels, has made significant strides.Every day, millions of status updates, posts, and Tweet messages are made and posted on websites like Facebook, Yammer, and Twitter to reflect people's current attitudes and opinions regarding specific agendas. However, due to the distinctive features that micro blogs like twitter contain, sentiment analysis of these blogs is thought to be a considerably harder challenge [3].

[1] *AKGEC GhaziabadDepartment name CSE AKGEC Ghaziabad U.PCity name, ynp1581@gmail.com*
[2]*Assistant Professor, School of Engineering, Department of CSE, Mallareddy University, drsesadri@mallareddyunivrsity.ac.in*
[3]*Associate Professor, Department of Computer Science and Engineering Jain(Deemed-to-be University), Bangalore, India t.mahesh@jainuniversity.ac.in*
[4]*Student, Data Science and Engineering, Birla Institute of Technology Pilani, Rajasthan, India shreyanth0810@gmail.com/2020sc04876@wilp.bits-pilani.ac.in*
[5]*Professor, Department of CSE RIMT University, Mandi, Gobindgarh, Punjab, India, ashishoberoi@rimt.ac.in*
[6]*Dr. A.P.J. Abdul Kalam Technical University, Uttar Pradesh Computer Science & Engineering, Dewan VS Institute of Engineering & Technology, Meerut, kpjayant@gmail.com, jayant@dewaninstitutes.org*

2. Literature review:

The literature reviews on sentiment analysis of SNS data for depression assessments are included in this part. In [4], the author suggested a multi-kernel SVM-based model to identify sad individuals and extracted three kinds of data from users' social media profiles, including user microblog content, user behaviours, and user profiles, to represent users' circumstances. The authors of [5] presented a data mining application based on categorization techniques—decision trees C 4.5 and J 4.8—to identify individuals who will likely experience depression in the future. To choose the optimal feature from the training data, the author of [6] employed a hybrid ML algorithm technique called a SVM. Marketing researchers rely on customers' recall of their felt experiences in surveys, which can be very variable and challenging to articulate and reconstruct [7]. In contrast, there are worries about the artificial conditions under which data is acquired in tests, which may limit customers' emotional responses [8]. Due to the enormous percentage of online interactions that convey consumers' thoughts, feelings, and opinions regarding products and brands, social media platforms are now widely used to analyse consumer sentiment on a broad scale and in a natural setting. According to [9], automated sentiment analysis is gaining more and more attention from both academia and business. It is becoming one of the most important methods for managing massive amounts of social media data [10].

3. Proposed social media data in sentimental analysis:

This section discuss novel technique in machine learning algorithm in big data for social media based sentimental analysis. Here the input data has been collected as social media based sentimental data and processed for noise removal. Then this data has been clustered using Fuzzy-C means clustering and feature extracted using differential multi-layer whale optimization. The overall proposed architecture is shown in figure-1.

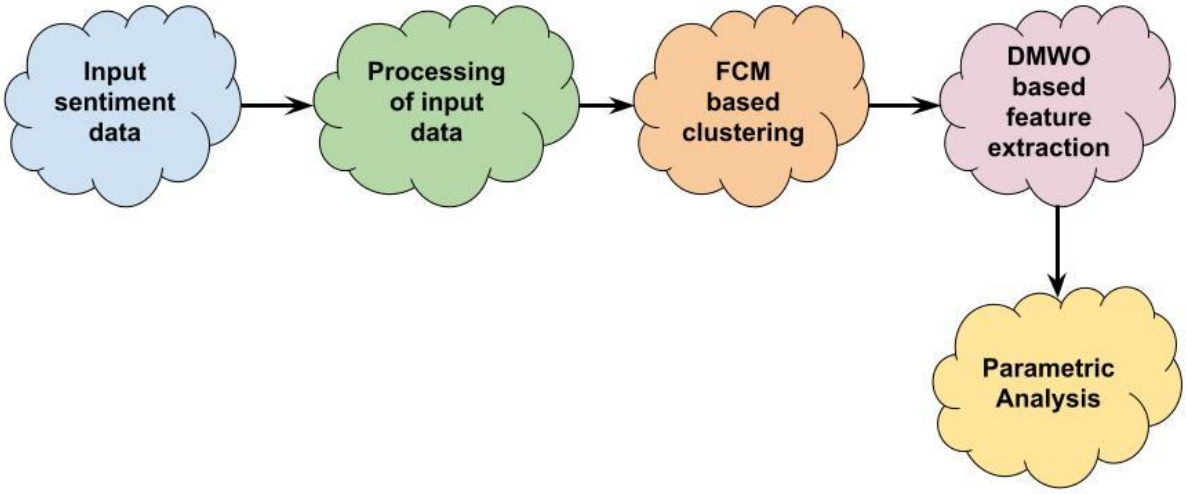

Figure 1. Overall proposed architecture

Raw analysis Usually, data scraped from the internet yields noisy data. This is a result of people using social media in a casual way. Many preparation techniques were used to normalise the dataset and reduce its size.

Fuzzy-C means clustering with differential multi-layer whale optimization:

Let us consider the dataset $Z = \{z_1, z_2, \dots, z_q\}$ with cluster set $X = \{x_1, x_2, \dots, x_p\}$ and membership set W= $\{w_{kl} \mid 1 \le k \le e, 1 \le l \le p\}$ further considering these three FCM can be formulated. The suggested mechanism's general premise is to combine an IFCM with a double neural network. To train the example, we further create an efficient auto-encoder by eq. (1).

$$\min: \sum_{k=1}^{e}\sum_{l=1}^{p} w_{kl}^{o}\|z_l - x_k\|^2 \quad \sum_{l=1}^{e} w_{kl} = 1, w_{kl} \ge 0 \qquad (1)$$

We create modified-FCM, or Modified FCM, in the equation (2) below to prevent spurious clustering.

$$L_o(W, X) = \sum_{k=1}^{e}\eta_i\sum_{k=1}^{p}(1 - u_{kl}^{o})^{\circ} + \sum_{k=1}^{e}\sum_{k=1}^{p} w_{kl}^{m}\|z_l - z_i\|^2 \qquad (2)$$

In order to update the membership matrix and cluster centres, the equation can be optimised as eq. (3):

$$x_k = \sum_{l=1}^{p} w_{kl}^{o} z_l / \sum_{l=1}^{p} w_{kl} \qquad (3)$$

Membership matrix by eq. (4)

$$w_{kl} = \left(1 + \left(\frac{e_{kl}}{\eta_k}\right)^{-1/(o-1)}\right)^{-1} \qquad (4)$$

The term e_{kl} in the equation above denotes the separation between the membership matrix and the cluster.

The donor vector provides a certain number of components to the target vector, denoted by the integer index L, so that L ∈ [1, D]. A trial vector is then calculated as eq. (5):

$$u_{z,j} = \begin{cases} v_{z,j} \text{ for } j = \langle l \rangle_D, \langle l+1 \rangle_D, \dots, \langle l+L-1 \rangle_D \\ x_{z,j} \forall j \in [1, D] \end{cases} \qquad (5)$$

Contrarily, the binomial crossover is applied to each variable with a known crossover probability, as eq. (6):

$$u_{z,j} = \begin{cases} v_{z,j} & \text{if (rand } \le cr \text{ or } j = j_{\text{rand}} \text{)} \\ x_{z,j} & \text{otherwise} \end{cases}$$

$$\alpha = 2\left(1 - \frac{t}{T}\right) \qquad (6)$$

where T stands for the greatest number of iterations and t stands for the current iteration. Generally speaking, expanding your search field will lessen your chances of experiencing local optima stagnation. Different strategies can be used to speed up exploration.In this regard, some non-linear functions are proposed to reduce the values of in order to balance the exploration and exploitation phases. These non-linear functions have varying slopes and distinct curve shapes by eq. (7).

$$\text{hid\_layer} \quad {}_{l_{11\dots},l_P} = \text{enc}(\psi)\left(\sum_{k_1\dots k_P}^{R_1\dots K_P} d_{l_1\dots l_P}^{(1)} + Y_{\alpha k_1\dots k_P}^{(1)} Z_{k_1\dots k_P}\right)$$

$$\text{out\_layer} \quad {}_{k_1,\dots k_P} = \text{dec}(\psi)\left(\sum_{l_1,\dots l_o}^{L_1,\cdots L_P} d_{k_1\dots\cdots-k_P}^{(1)}\right) \qquad (7)$$

In the equation above, K1 stands for the number of dimensions, L1 for the hidden layer, enc for encoder, and dec for decoder; also, the sigmoid function is used in both the encoding layer and the decoding layer.

4. Performance analysis:

Jack Dorsey may rapidly publish text, images, or videos via Twitter, which has a character limit of 280. Additionally, you can follow other accounts, enjoy tweets from other accounts, or retweet them (Rogers, 2014).For this study, the Twitter (Application Programming Interface) API was used to gather 4500 health-related tweets. A Python application was used to preprocess the data and calculate sentiment scores.1680 of the tweets that were gathered and classified as neutral, 1220 as positive, and 1600 as negative.The same models were also applied to the dataset of 500 good and 500 negative opinions compiled from the IMDB movie reviews in addition to the study on the data gathered from Twitter.The Twitter data gathered revealed that the medication advertisements were the neutrally labelled messages. Tweets with a negative rating appear to come from people who have different conditions.On the other hand, the tweets showing that diseases like cancer have been effectively cured are the positive ones.

Table-1 Comparative analysis between proposed and existing technique

Parameters	SVM	ASA	MLA_BDA_SMDSA
Accuracy	89	93	95
Precision	64	69	72
Recall	55	59	62
AUC	41	43	44
RMSE	44	49	52

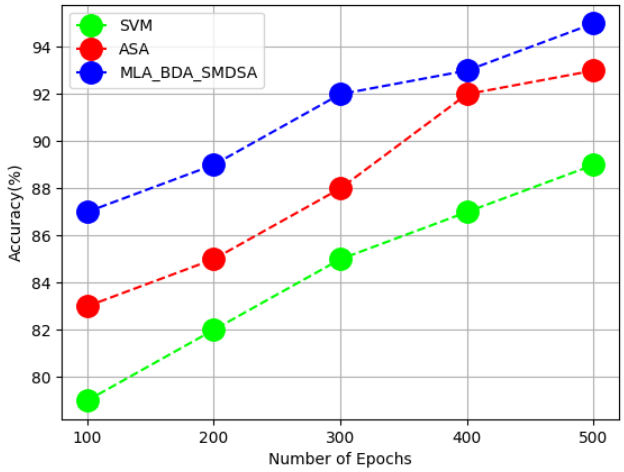

Figure-2 Comparison of accuracy

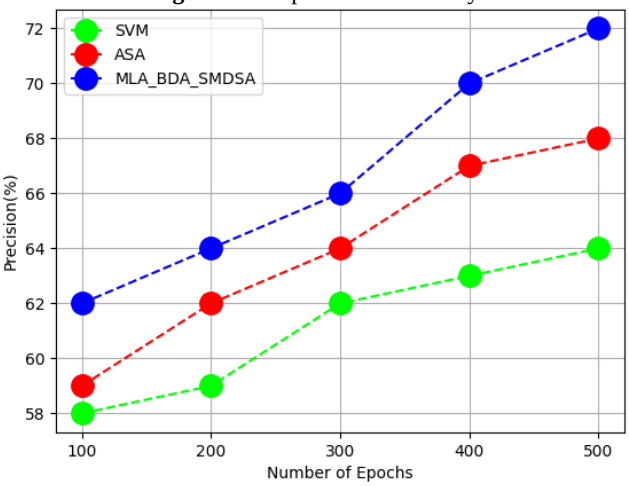

Figure-3 Comparison of precision

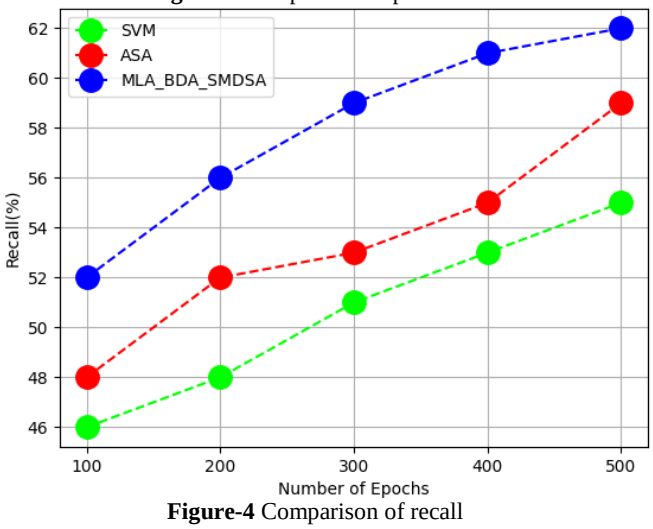

Figure-4 Comparison of recall

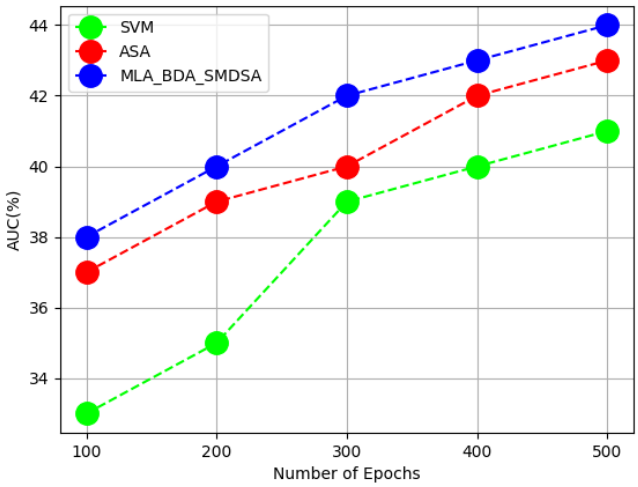

Figure-5 Comparison of AUC

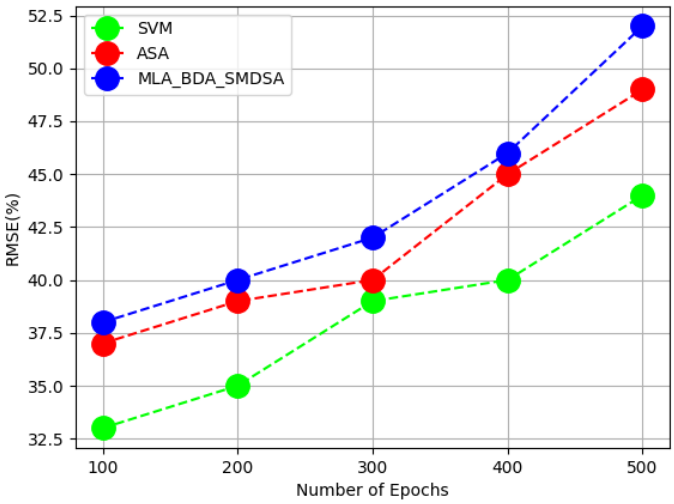

Figure-6 Comparison of RMSE

The above table-1 shows comparative analysis between proposed and existing technique for social media data. Here the parametric analysis has been carried out in terms of accuracy, precision, recall, AUC and RMSE. The proposed technique attained accuracy of 95%, precision of 72%, recall of 62%, AUC of 44%, RMSE of 52% as shown in figure 2-6.

5. Conclusion:

This research proposenovel technique in machine learning algorithm in big data for social media based sentimental analysis. The processed input data has been clustered using Fuzzy-C means clustering and feature extracted using differential multi-layer whale optimization. Sentiment analysis looks for and extracts opinions and attitudes regarding a particular topic from a given piece of text.This sentiment analysis technique creates an artificial intelligence that is smarter and more human-like and can assess and reply in a particular way based on the emotions users display in a textual chat, tweet, or blog conversation. The proposed technique attained accuracy of 95%, precision of 72%, recall of 62%, AUC of 44%, RMSE of 52%..

References

[1] Jindal, K., & Aron, R. (2021). A systematic study of sentiment analysis for social media data. *Materials today: proceedings*.

[2] Behera, R. K., Jena, M., Rath, S. K., &Misra, S. (2021). Co-LSTM: Convolutional LSTM model for sentiment analysis in social big data. *Information Processing & Management*, *58*(1), 102435.

[3] Hansen, K. B., &Borch, C. (2022). Alternative data and sentiment analysis: Prospecting non-standard data in machine learning-driven finance. *Big Data & Society*, *9*(1), 20539517211070701.

[4] Biradar, S. H., Gorabal, J. V., & Gupta, G. (2022). Machine learning tool for exploring sentiment analysis on twitter data. *Materials Today: Proceedings*, *56*, 1927-1934.

[5] Alharbi, M. S., & El-kenawy, E. S. M. (2021). Optimize machine learning programming algorithms for sentiment analysis in social media. *International Journal of Computer Applications*, *174*(25), 38-43.

[6] Kaur, H., Ahsaan, S. U., Alankar, B., & Chang, V. (2021). A proposed sentiment analysis deep learning algorithm for analyzing COVID-19 tweets. *Information Systems Frontiers*, *23*(6), 1417-1429.

[7] Babu, N. V., &Kanaga, E. (2022). Sentiment analysis in social media data for depression detection using artificial intelligence: A review. *SN Computer Science*, *3*(1), 1-20.

[8] Ahmed, H. M., Javed Awan, M., Khan, N. S., Yasin, A., & Faisal Shehzad, H. M. (2021). Sentiment analysis of online food reviews using big data analytics. *Hafiz Muhammad Ahmed, Mazhar Javed Awan, Nabeel Sabir Khan, Awais Yasin, Hafiz Muhammad Faisal Shehzad (2021) Sentiment Analysis of Online Food Reviews using Big Data Analytics. Elementary Education Online*, *20*(2), 827-836.

[9] Alsayat, A. (2022). Improving Sentiment Analysis for Social Media Applications Using an Ensemble Deep Learning Language Model. *Arabian Journal for Science and Engineering*, *47*(2), 2499-2511.

[10] Mishra, R. K., Urolagin, S., Jothi, J. A., Neogi, A. S., & Nawaz, N. (2021). Deep learning-based sentiment analysis and topic modeling on tourism during Covid-19 pandemic. *Frontiers in Computer Science*, *3*(10.3389).

.