Journal of Engineering and Sustainable Development
Vol. 28, No. 05, September 2024
https://jeasd.uomustansiriyah.edu.iq/index.php/jeasd
ISSN 2520-0917
Research Article
https://doi.org/10.31272/jeasd.28.5.13

A Hybrid for Analyzing Text Streaming Using Data Mining and Machine Learning Techniques

Maysaa Hameed Abdulameer[1], Mahmood Zaki Abdullah[2*], Ali Khalid Jassim[3], Mohammed Majid M. Al Khalidy[4]

[1]Iraqi Commission for Computer & Informatics (ICCI), Informatics Institute for Post-Graduation Studies (IIPS), Baghdad, Iraq
[2]Computer Engineering Department, College of Engineering, Mustansiriyah University, Baghdad, Iraq
[3]Electrical Engineering Department, College of Engineering, Mustansiriyah University, Baghdad, Iraq
[4]Department of Electrical and Electronics, College of Engineering, University of Bahrain, Manama, Kingdom of Bahrain

[1]https://orcid.org/0009-0000-7182-4275
[2]https://orcid.org/0000-0002-3191-3780
[3]https://orcid.org/0000-0002-4146-4536
[4]https://orcid.org/0000-0002-9723-0405
[*]Email: drmzaali@uomustansiriyah.edu.iq

Article Info

Received	04/11/2023
Revised	14/08/2024
Accepted	17/08/2024

Abstract

Human opinions and feelings can be studied and analyzed in various fields. Sentiment analysis divides data into neutral, positive, and negative categories to classify a writer's or speaker's attitude toward various topics or events. This study uses a hybrid approach that combines (Particle_Swarm_Optimization PSO) with machine learning classifiers (Artificial Neural_ Networks ANN, Naïve Base NB, and Support.Vector.Machine SVM). Following the preprocessing phase of each data collection, a Convolution Neural Network (CNN) will be used to create feature vectors, preparing the raw data as a stored labeled dataset with three classes (positive, negative, and natural). After that, with two hidden layers whose parameters were optimized by PSO, this dataset will be prepared for classification using NB, SVM, and ANN. The best results will be obtained with the proposed Artificial Neural Network A.N.N which is close to 99% Acc. value, about SVM Acc. equal to 92%, and NB 89%, better than using the same ML algorithms.

Keywords: Machine Learning, Deep Learning, Convolution-Neural-Network, Practical-Swarm-Optimization

1. Introduction

Dynamic data are continuously created data streams. This makes it possible to examine data in real time and obtain an understanding of a variety of situations. Sentiment analysis on social media (SM) is one application of data streams [1].

Because S.M. is incorporated in many data sources at various dim., it is frequently reported to be rather vast. SM promotes commercial operations and is fundamental to human communication [2]. A well-known social media app has been displayed in Fig.1, which consumers have been using more and more in recent years. From tweets to swipes, likes to shares, these social media platforms create a massive amount of data; Terabytes to Petabytes of data are being created online every day [3], [4]. There are currently over a billion more internet users than there were just a few years ago. Because they have comparatively easy access to social media, cell phones have a share to growth in nearly half of the world's online-traffic. To reap the rewards of machine learning and deep learning, SM also uses them, Fig. 2 illustrates the general processing steps of SM through machine learning [2],[5].

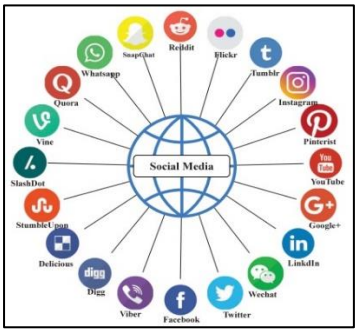

Figure 1. Common Social Media Platforms [3]

The field of networks has a plethora of SM applications. SM like Facebook, Snapchat, QQ, WhatsApp, WeChat, QZone,

Tumblr, Instagram, Twitter, and so on applications that are used daily [6], [7]. Fig. 3 illustrates how machine learning is applied to social media analysis through machine learning

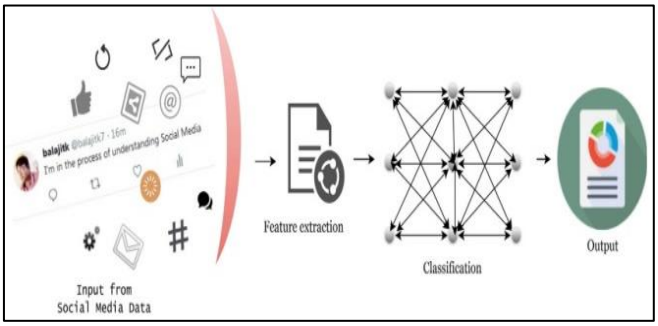

Figure 2. Social media through Machine Learning [2]

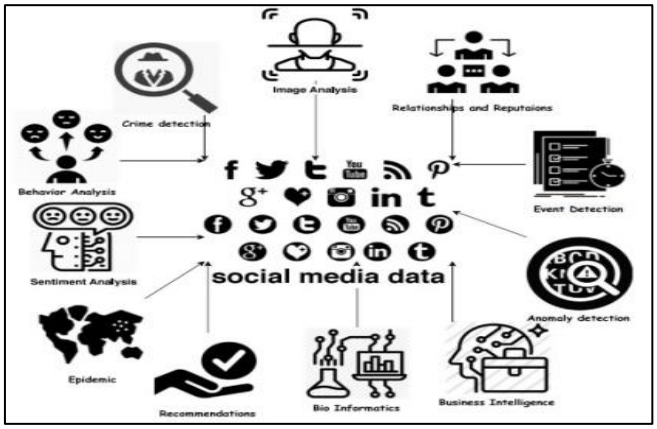

Figure 3. Applications of SM analysis through ML

While there are many machine learning approaches available, only a few numbers are used in the context of social network analytics. Additionally, to increase the output's efficiency, several social analytics methodologies require adjustments to the learning method or the environment [8], [9].

A textual data set's opinions or feelings are analyzed through a process known as sentiment analysis or opinion mining. This text analysis technique finds polarity in any text, sentence, paragraph, or clause that an individual posts about a particular event. Intentions, such as being interested or not, are also highlighted, along with feelings and emotions (such as happiness, sadness, anger, etc.) [10]. Within the context of social network mining and multimedia data mining, sentiment analysis is one of the main research topics. As in mining, it is the process of extracting and discovering patterns and knowledge from large datasets. It's like sifting through a giant pile of dirt to find the hidden gems, or in this case, valuable insights. Research endeavors can derive significant insights into sentiment analysis by utilizing this multimedia effectively. When paired with social multimedia, sentiment analysis facilitates the analysis of people's opinions, a range of emotions, and attitudes toward data written in textual languages [11], several types of Sentiment analysis such as Fine-grained sentiment analysis, Emotion detection, Aspect based sentiment analysis, and Multi-lingual sentiment analysis [12], [13]. In this study, we will concentrate on utilizing a social media application called sentiment analysis, which calls for the use of an optimization technique called PSO in conjunction with a hybrid neural network and machine learning approach called Artificial Neural-Deep Neural Network (ANN-DNN) learning.

2. Related Work

Analyzing the airline industry's services on Twitter, Sreenivasulu et al. [14] focus on using the intelligent deep neural network (IDNN), which is caused by fuzzy and has access to chaotic PSO information. Through the integration of chaotic PSO, the fuzzy neural network is optimized. The rendering of the developed I.D.N.N is compared with the random forest and SVM classifiers. Classifiers and PSO algorithms are combined in a hybrid approach used by Badr et al. [15], for every tweet, preprocessing is done before the binary text categorization. After that, the TF-IDF is calculated to create their feature vectors, and (ACO) is used for optimization. NB and SVM were utilized for tweet binary classification. Using preprocessed tweets to gather feedback or opinions from their respective passengers through sentiment analysis Al-Amrani et al. [16] proposed a deep-belief model for sentiment classification in their work. This model usefully classifies sentiment by leveraging its prior knowledge of the deep structure of the network. Devikanniga et al. [17] created a sentiment analysis from product reviews using a hybrid technique. The model classifies the product review as either pos. or neg. based on the question posed in the relevant product reviews that Amazon offers. To improve classification accuracy, Rathi et al. [18] present a hybrid model based on SVM and KNN; however, most of the literature in this field relates to 2_way classification. Naiknaware et al. [19] show the classifiers' performance after they were developed for seven datasets. In the Budget2017 dataset, NB performs the best. In the GST 2017, SVM yielded the best results; the top-performing regions were made in India, Digital India, Kashmir, and Startup Max Entropy. They also find that the Mean Absolute Error can be easily predicted using the Mean Error. Kameswara et al. [20] The purpose of this paper is to dissect printed content from Facebook. The data that was gathered was integrated with special content mining systems. The proposed work intends to explain the model of user discretion about a subject using R Studio.

3. The Proposed Methodology

The following is the suggested approach for sentiment analysis utilizing a hybrid neural network (ANN-DNN) and machine learning, with PSO used to adjust the classifiers' parameters for optimal performance, for collecting the data the API platform has been used. Data can be gathered from Twitter and Facebook pages. Each of these platforms has billions of users who write and comment on their posts every week. Following text data collection, the data is processed to remove any null values and outliers, by sampling the data using appropriate techniques. The majority of the data that was retrieved includes messages along with time stamps, URLs, emoticons, special characters, stop words, emoticons, and hashtags. Thus, the data will be processed to prepare for mining, the processing includes (Translation, Converting, Removal of stop words ($, @, #, %), Tokenizing, and stemming the words into root words. Two

main processes of the suggested approach which explain in the following:

- Feature Extraction: several feature extraction methods are based on CNN, The data will be processed to extract features using this model. The network's convolutional, pooling, and fully connected components are its essential elements. The input collected data is passed through several conv. kernels in the conv. layer to create the feature map, a set of weights known as the kernel is multiplied by the input in the linear process known as convolution. The dot product is then summed, and this always yields a single value from top to bottom and left to right. In the pooling layer, the feature map produced by the previous layer's convolution operation is down-sampled. The network becomes more complex as the convolutional and pooling layers are iterated. Fig. 4 describes how the suggested CNN is structured in terms of feature extraction:

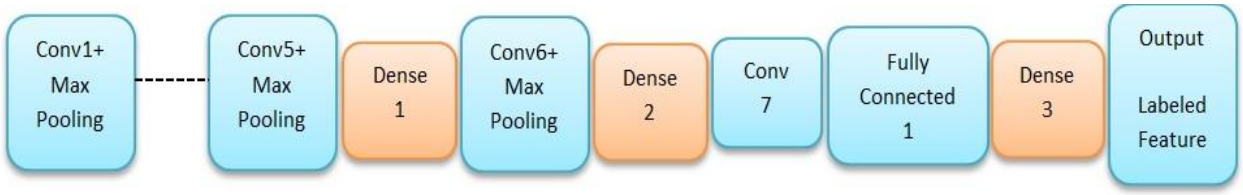

Figure 4. The Proposed CNN Model

Table 1 explains the layers and the number of parameters used in the proposed CNN network:

Table 1. The Layers and Parameters of the CNN

Layer (type)	Output Shape	Parameters
Conv.1 (Conv1D)	(No., 7498, 16)	64
Max-pooling. 1	(No., 7498, 16)	0
Conv.2 (Conv1D)	(No., 7496, 32)	1568
Max-pooling. 2	(No., 7496, 32)	0
Conv.3 (Conv1D)	(No., 7494, 32)	3104
Max-pooling. 3	(No., 7494, 32)	0
Conv.4 (Conv1D)	(No., 7492, 128)	12416
Max-pooling. 4	(No., 7492 128)	0
Conv.5 (Conv1D)	(No., 7490, 128)	49280
Max-pooling. 5	(No., 7490, 128)	0
Dense.1 (Dense)	(No., 7490, 128)	16512
Conv.6 (Conv1D)	(No., 7488, 64)	24640
Max-pooling. 6	(No., 7488, 64)	0
Dense.2 (Dense)	(No., 7488, 128)	8320
Conv.7 (Conv1D)	(No., 7488, 40)	5160
Flatten.1	(No., 299520)	0

- Classification: To obtain the best classification results, the parameters of each machine learning model will be adjusted using PSO. The data that has been labeled needs to be classified using a neural network with two hidden layers and two machine learning models (NB and SVM) with three classes: Normal, Positive, and Negative. The data will be split (60:40) this split gave the best results compared to the (70:30) or (80:20). Three random parts are taken from the dataset, and each part is further divided into three parts:
  - Train (60%): This is the input that the neural network uses to learn.
  - Test (20%): This determines whether or not the suggested NN is overfitting.
  - Validation (20%): This part evaluates how well the algorithm is being executed.

The (60:40) splitting is the best result than splitting (70:30) or (80:20) when the data was split in the paper because this strongly supports the research results. where most papers have their best results when they are (80:20).

The sigmoid function was used in the proposed neural network layers. Since each hidden node is described by "Eq. “Eq. (1)”:

$$(x)=(\boldsymbol{P_i^T}x+\theta i) \quad (1)$$

For a neural network with H hidden layers, the entire equation is as in “Eq. (2)”:

$$(x)=\sum_{i=1}^{H} \boldsymbol{vioi}(\boldsymbol{x}) \quad (2)$$

The neural network's training error may become stuck in local minima as a result of the sign. func. being 1 or 0. The injection of the PSO method is used in the suggested method to limit the parameters of the neural network to a range of values, the best epoch number used is 100 epochs. Using L as a limit for the absolute value. of the para. x of the sigmoid function, the quantity Q(L) is estimated. In algorithm 1, the steps for this computation are displayed.

Algorithm (1) calculates the quantity Q(L), and M the number of patterns (x).

Step 1: Function Q(L)

Step2: Define v=0

Step 3: For i=1to$H$ Do

For j=1to$M$ Do

If $(|\sum_{k=1}^{d}\left(W_{(d+2)i-(d+i)+kX_{jk}}\right)+W_{(d+2)i}>L$ then v=v+1

End For j

Step 4: End For i

Step 5: Return $\frac{\boldsymbol{v}}{\boldsymbol{H*M}}$

Step 6: End

Artificial neural networks are trained using the PSO technique, which minimizes the error function and a penalty factor based on the function. (L) specified in algorithm1. Therefore, "Eq. (3)" will be minimized by the PSO technique:

$$\boldsymbol{E_T(N(x,w),L) = \sum_{i=1}^{M}(N(x_i,w)-y_i)^2 * \left(1+\propto B(L)\right)} \quad (3)$$

Algorithm 2 The P.S.O algorithm in one processing unit.

Step 1: Start

Set K=0, H, m * K the iteration num.

Set the max. numb. of iteration.

Set velocities. $u1,2,\ldots,um$ rand.

For i=1to$m$ do $pi,=pi$. The vector pi,b

Set $pbest$=arg. min $i\in1\ldots(pi)$

Step 2: If $k\geq kmax$, then Stop.

Step 3: For i=1to$m$ Do

Compute the velocity ui using the vectors $ui,pi.b$, and $pbest$

Set the new position $pi=pi+ui$

Calculate the $f(pi)$ using $f(pi)=ET(N(x,pi),L)$

If$f(pi)\leq f(pi,b)$ then $pi,b=xi$

Step 4: End-For

Step 5: Set $pbest$=arg. Min. $i\in1\ldots(pi)$

Step 6: Set $k=k+1$.

Step 7: Go to Step 2

Algorithm 2 is the PSO algorithm as above:
"Eq. (4)" is typically used to calculate each particle's velocity:

$$ui = \omega ui + r1c1(pi - xi) + r2c2(pbest - xi) \quad (4)$$

where
$r1,2$ define randomly numbers in [0,1], $c1,2$ are constants, ω is inertia, the inertia calculation is defined as in "Eq. (5)"

$$\omega k=\frac{\text{kmax}-\text{k}}{\text{kmax}}(\omega\text{max}-\omega\text{min})+\omega\text{min} \quad (5)$$

where ωmin and ωmax represent the inertia's minimum and maximum values, respectively. The vector $pbest$ stores the ideal set of parameters after the PSO is finished. It is possible to begin a local optimization technique from this set to reduce the neural network error even further. Additionally, an interval for the neural network's para.(s). can be computed using the ideal set of weights. During this time, the error function will be reduced to the minimum. The following procedures are used to calculate the interval $[LW]$ for the neural network's parameter vector w:

1. **For** i=1…n d**o**
   Set $LWi=-F\times|pbest|$
   Set $RWi=F\times|pbest|$
2. **End.For**

The value F will be called the margin factor with F>1.

4. Performance Evaluation

The three classifiers must be evaluated to determine which PSO training method produced the highest accuracy, without regard to the efficacy of the optimization technique used to tweak the classifiers' parameters.

There are four basic terms for evaluating machine learning algorithms:

1. True-Positives (TP): the actual result matched our prediction of "yes."
2. True-Negatives (TN): the actual result matched our prediction of "No."
3. False-Positives (FP): This is when we expected a certain outcome, but it turned out to be false.
4. False-Negatives (FN): This is when we expected a certain outcome, but it turned out to be true.

Other metrics exist that could influence accuracy: According to the equations "Eq. (6), Eq. (7), Eq. (8), Eq. (9)" below [21], [22], precision, recall, and F score:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (6)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (7)$$

$$\text{F Score} = 2 * \frac{(\text{Recall}*\text{Precision})}{(\text{Recall+Precision})} \quad (8)$$

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{Tp+TN+FP+FN}} \quad (9)$$

5. Results and Discussions

The classification stage results, both with and without the use of the PSO optimization method, will be examined in this section first. Before that, Fig. 5 illustrates the true class counts of positive, negative, and neutral for the tweet and Facebook data. The positive number is 4363, the neutral number is 6088, and the negative number is 17175. The hybrid ANN-DNN learning model that is being proposed makes use of a PSO technique along with multiple thin layers.

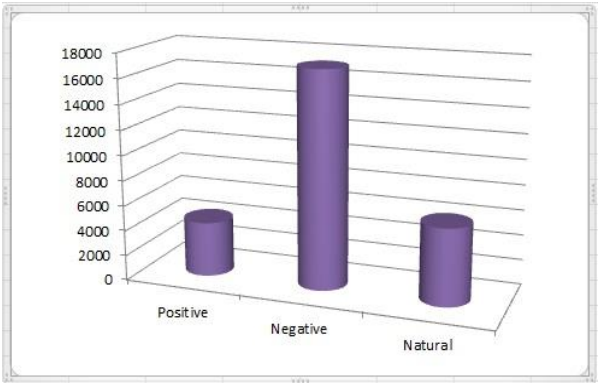

Figure 5. The Actual Count of Dataset Classes

The accuracy outcomes of the three machine learning algorithms with PSO are ANN = 99 %, SVM=92%, and NB=89%. Fig. 6 shows the accuracy results of each machine learning.

The suggested approach contrasts the sentiment analysis findings of data gathered from Facebook and Twitter apps, as shown in Table 2 and Fig 7 to Fig. 9 which represent the machine learning outcome metrics using the PSO optimization. Through the results in Table 2, it becomes clear to us the importance of using the PSO optimization algorithm with machine learning classification methods.

The PSO algorithm generally increased the research accuracy results for all three machine learning algorithms, according to the results of all machine learning. The above figures illustrate the degree to which employing an optimization technique has an impact.

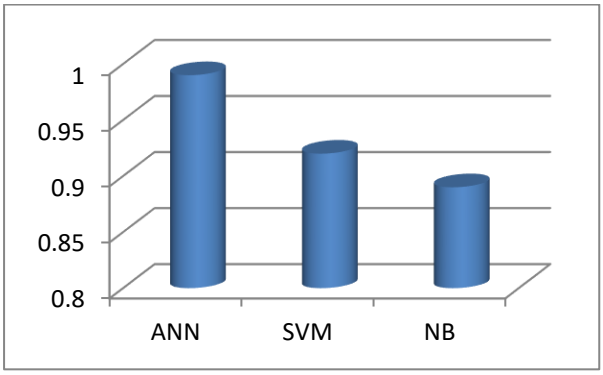

Figure 6. The Accuracy Results of Machine Learning

Table 2. The Performance Measures

Algorithms/Measures	Precision	Recall	F1 SCORE	Accuracy
ANN-DNN with PSO	0.99	0.98	0.97	0.99
ANN-DNN without PSO	0.87	0.85	0.85	0.88
SVM with PSO	0.85	0.74	0.84	0.92
SVM without PSO	0.80	0.77	0.80	0.80
NB with PSO	0.85	0.90	0.82	0.89
NB without PSO	0.71	0.87	0.78	0.78

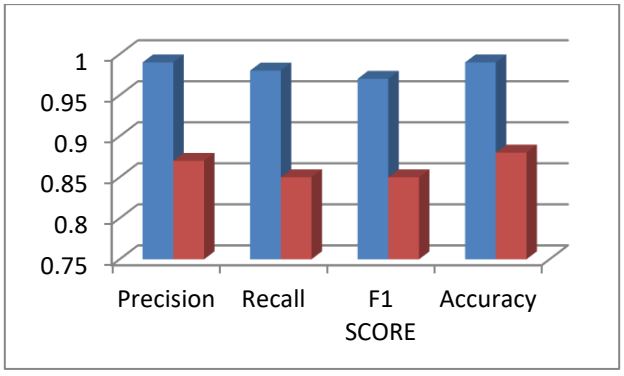

Figure 7. Results of the Metrics of ANN

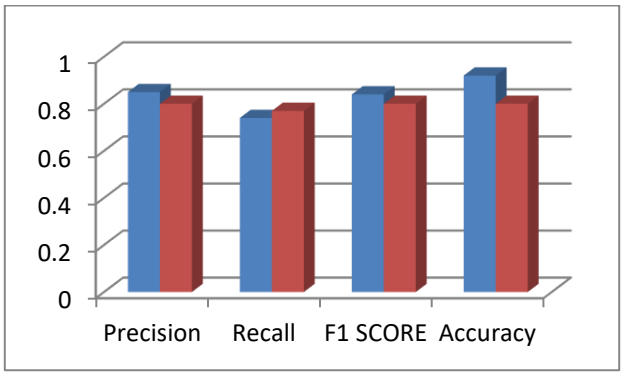

Figure 8. Results of the Metrics of SVM

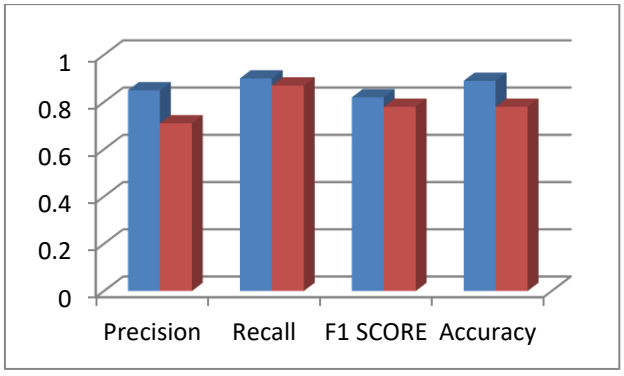

Figure 9. Results of the Metrics of NB

6. Conclusion

Sentiment analysis is a technique for figuring out people's attitudes, beliefs, and feelings. Men and women alike can naturally hold negative or positive opinions. Speech elements are widely used in textual content analysis to extract sentiment. It can be difficult to tell the difference between sentiment and opinion when adjectives and adverbs are used together. This work established, examined, and evaluated hybrid neural networks and machine learning for the PSO-optimized Sentiment Analysis task. From the experiment, we learned a great deal about how to deal with machine learning, deep learning, neural networks, and data analysis issues as well as how to analyze data to facilitate the learning process for machine learning algorithms. Text classification difficulty is one of the most important things to keep in mind when working

with the kinds of text and phr. is to be in the data. This is because it has a substantial impact on the wide range of words that machine learning can identify.

Conflict of interest

The authors ensure that there is no conflict in publishing this paper.

Author Contribution Statement

Maysaa Hameed Abdulameer proposed the aim and algorithms of this paper, while Mahmood Zaki Abdullah, Ali Khalid Jasim, and Mohammed Majid M. Al Khalidy participated in making calculations and verifying the algorithms.

References

[1] T. Kolajo, O. Daramola, and A. Adebiyi, "Big data stream analysis: a systematic literature review," *Journal of Big Data*, vol. 6, no. 1, Jun. 2019, doi: https://doi.org/10.1186/s40537-019-0210-7.

[2] B. T.K., C. S. R. Annavarapu, and A. Bablani, "Machine learning algorithms for social media analysis: A survey," *Computer Science Review*, vol. 40, no. 1, p. 100395, May 2021, doi: https://doi.org/10.1016/j.cosrev.2021.100395.

[3] T. Sravanthi, V. Hema, S. Tharun Reddy, K. Mahender, and S. Venkateshwarlu, "Detection of Mentally Distressed Social Media Profiles Using Machine Learning Techniques," *IOP Conference Series: Materials Science and Engineering*, vol. 981, p. 022056, Dec. 2020, doi: https://doi.org/10.1088/1757-899x/981/2/022056.

[4] D. Anashkin and K. Malyshenko, "Social Media Content Analysis with Machine Learning Tools." Accessed: Aug. 13, 2024. [Online]. Available: https://ceur-ws.org/Vol-2914/paper1.pdf

[5] Ferda Ofli, F. Alam, and M. Imran, "Analysis of Social Media Data using Multimodal Deep Learning for Disaster Response," *arXiv (Cornell University)*, Jan. 2020, doi: https://doi.org/10.48550/arxiv.2004.11838.

[6] A. Goswami *et al.*, "Sentiment Analysis of Statements on Social Media and Electronic Media Using Machine and Deep Learning Classifiers," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–18, Mar. 2022, doi: https://doi.org/10.1155/2022/9194031.

[7] H. Taherdoost and M. Madanchian, "Artificial Intelligence and Sentiment Analysis: A Review in Competitive Research," *Computers*, vol. 12, no. 2, p. 37, Feb. 2023, doi: https://doi.org/10.3390/computers12020037.

[8] A. Patil, "A Survey on Machine Learning Approaches to Social Media Analytics," *www.academia.edu*, vol. 11, no. 4, 2016, Accessed: Dec. 19, 2022. [Online]. Available: https://www.academia.edu/34842090/A_Survey_on_Machine_Learning_Approaches_to_Social_Media_Analytics

[9] A. Capatina, M. Kachour, J. Lichy, A. Micu, A.-E. Micu, and F. Codignola, "Matching the future capabilities of an artificial intelligence-based software for social media marketing with potential users' expectations," *Technological Forecasting and Social Change*, vol. 151, no. 151, p. 119794, Feb. 2020, doi: https://doi.org/10.1016/j.techfore.2019.119794.

[10] S. Bhushan and S. Shanker, "A Review On Machine Learning Techniques On Social Media Data For Policy Making," *International Research Journal of Engineering and Technology (IRJET)*, vol. 7, no. 6, 2020.

[11] M. Rashid, A. Hamid, and S. A. Parah, "Analysis of Streaming Data Using Big Data and Hybrid Machine Learning Approach," *Handbook of Multimedia Information Security: Techniques and Applications*, pp. 629–643, 2019, doi: https://doi.org/10.1007/978-3-030-15887-3_30.

[12] A. P. Rodrigues *et al.*, "Real-Time Twitter Spam Detection and Sentiment Analysis using Machine Learning and Deep Learning Techniques," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–14, Apr. 2022, doi: https://doi.org/10.1155/2022/5211949.

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

[14] T. Sreenivasulu and R. Jayakarthik, "Intelligent Deep Neural Network Integrated with Chaotic Particle Swarm Intelligence based Sentiment Analysis in Big Data Paradigm," *7th International Conference on Advanced Computing & Communication Systems (ICACCS)*, Mar. 2021, doi: https://doi.org/10.1109/icaccs51430.2021.9441976.

[15] E. M. Badr, M. Abdul, M. Ali, and H. Ahmed, "Social Media Sentiment Analysis using Machine Learning and Optimization Techniques," *International Journal of Computer Applications*, vol. 178, no. 41, pp. 31–36, Aug. 2019, doi: https://doi.org/10.5120/ijca2019919306.

[16] Y. Al Amrani, M. Lazaar, and K. E. El Kadiri, "Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis," *Procedia Computer Science*, vol. 127, pp. 511–520, 2018, doi: https://doi.org/10.1016/j.procs.2018.01.150.

[17] D. Devikanniga, A. Ramu, and A. Haldorai, "Efficient Diagnosis of Liver Disease using Support Vector Machine Optimized with Crows Search Algorithm," *EAI Endorsed Transactions on Energy Web*, vol. 0, no. 0, p. 164177, Jul. 2018, doi: https://doi.org/10.4108/eai.13-7-2018.164177.

[18] M. Rathi, A. Malik, D. Varshney, R. Sharma, and S. Mendiratta, "Sentiment Analysis of Tweets Using Machine Learning Approach," *2018 Eleventh International Conference on Contemporary Computing (IC3)*, vol. 6, no. 4, Aug. 2018, doi: https://doi.org/10.1109/ic3.2018.8530517.

[19] B. Naiknaware, B. Kushwaha, and S. Kawathekar, "Social Media Sentiment Analysis using Machine Learning Classifiers," *International Journal of Computer Science and Mobile Computing*, vol. 6, no. 6, pp. 465–472., 2017.

[20] R. M., Kameswara and D. G. Chandini, "Social Media Analytics Using Machine Learning," *International Journal of Pure and Applied Mathematics*, vol. 16, no. 6, 2017.

[21] R. Sawhney, P. Manchanda, P. Mathur, R. Shah, and R. Singh, "Exploring and Learning Suicidal Ideation Connotations on Social Media with Deep Learning," *ACLWeb*, Oct. 01, 2018. https://aclanthology.org/W18-6223/

[22] A. Nistor and E. Zadobrischi, "The Influence of Fake News on Social Media: Analysis and Verification of Web Content during the COVID-19 Pandemic by Advanced Machine Learning Methods and Natural Language Processing," *Sustainability*, vol. 14, no. 17, p. 10466, Aug. 2022, doi: https://doi.org/10.3390/su141710466.