Hate Speech Detection on Multiple Social Networks Using Deep Learning and Optimization Techniques: A Hybrid Approach

Vishu Tyagi, Sourabh Jain
Department of Computer Science and Engineering, Indian Institute of Information Technology, Sonepat, Haryana, India

***Abstract*—The use of social media networks as a source of hate speech is another emerging factor that complicates the possibility of a comprehensive organization of an environment suitable for promoting healthy communication. Automating the detection of hate speech in various social media networks has turned out to be a very difficult process. It is critical to identify and monitor hate speech to reduce its negative effects on people and groups. Currently, there are many approaches to classifying hate speech, but they still have indeterminacy when it comes to distinguishing between hate and normal messages and low accuracy. Many domains have greatly benefited from deep learning, especially in speech and NLP tasks. The hyperparameters of Deep Neural Networks (DNN) play a crucial role and are reflected in their success. However, because these hyperparameters are highly recursive, it is sometimes difficult to set them for machine learning models, such as deep neural networks. The work proposed in this study employed the sparrow search algorithm (SSA) optimization methods to fine-tune the hyperparameters of deep learning models for hate speech detection. In the training process of the SSA-DNN model, the SSA can help search and select the best hyperparameters. Based on the obtained experimental outcomes, it can be observed that the proposed SSA-DNN model outperforms different machine learning and deep learning techniques in the context of hate speech detection.**

Keywords—Natural language processing; sparrow search algorithm; hate speech; deep neural network; social media

I. Introduction

Most internet users engage with social media and microblogging platforms, which have become dynamic spaces for public discourse and discussion. Social media networks that were relatively popular in the past, including Twitter, YouTube, Facebook, Wikipedia, Gab, and Reddit, are now used by people of various ages and interests. These platforms detect rapid content growth, leading to the emergence of what is known as "big data." The explosion of data has created a lot of enthusiasm for researchers, as well as those involved in automated opinion analysis and the study of both the structural aspects of networks and how users communicate with one another. Although developed to provide an open forum for communication and support free expression, these systems are very difficult to moderate because of their massive quantity of communications and unique attributes. Further complicating matters is that the vastly different social, emotional, and cultural backgrounds of users on these sites can lead to hate speech and other forms of offensive language being expressed when users interact with each other in a way they do not agree upon.

Artificial neural networks (ANNs) are widely recognized as a foundational form of deep neural networks applied to various natural language processing (NLP) tasks [1], including the detection of hate speech. Deep neural networks (DNNs) have been developed for pattern recognition and are highly valued in academia because of their strong capability to model complex nonlinear relationships. Nevertheless, when trained on limited datasets, the performance of deep learning models often falls below that of simpler neural network architectures. Designing an optimized model is a challenging and time-consuming task because it requires the careful selection of a suitable network architecture along with appropriate hyperparameters [2].

Hyperparameter selection for a Deep Neural Network is primarily responsible for the performance of the Deep Neural Network. Hyperparameter Optimization (HPO) is necessary to ensure the success of Deep Learning solutions in all areas of application. Most traditional methods used today (manual tuning, Grid Search, Random Search), while common, are limited in their ability to find the best hyperparameters in large dimensional spaces, are computationally expensive, and provide little insight into which hyperparameters are chosen to be the "best" [3]. Therefore, there is a significant need for a method to efficiently and effectively identify the optimal configuration(s) of model architecture using HPO for Deep Learning Models. The process of HPO is typically described as a 'black box' procedure because finding the best combination of architectural hyperparameters requires the use of advanced optimization algorithms [4].

The rapid advancement of technology has significantly enhanced the collection, repository, and data processing of scientific research. Deep learning, which emulates the functioning of the human brain, has shown significant potential for data analysis owing to its exceptional learning capabilities, adaptability, and flexibility. It involves an Artificial Neural Network (ANN) with multiple layers that performs different data modifications [5]-[6]. However, designing a traditional deep learning approach involves defining the model's structure and setting hyperparameters. While a deep neural network, it has the following constraints when applied to time-series scenarios:

1) Deep Learning models need help in explaining their decisions and the parameters used for predictions.

2) Hyperparameters of deep neural networks are usually decided based on earlier analyses or observed learning details.

3) The performance of a Deep Neural Network (DNN) can be significantly enhanced by fine-tuning the hyperparameters, affecting the model's architecture and generalization.

Hence, there is a strong interest in reducing the impact of human features and finding the optimal hyperparameters. This study presents a deep learning approach [1],[7]-[10] to address these challenges using a Sparrow Search Algorithm [3] to optimize the parameters. The SSA approach helps resolve the issue of the deep learning network's inability to explain its learning process and mitigate human tendencies. This research aims to introduce a novel approach for hyperparameter optimization, showcasing outstanding performance in hate speech detection.

The key contributions of the study are described as follows:

- Proposes a novel SSA-DNN hybrid model for automated hyperparameter optimization in hate speech detection.
- Effectively optimizes key DNN parameters using SSA to improve accuracy and generalization.
- Evaluates the model on a large multi-source social media dataset for robust performance.
- Achieves superior accuracy compared to existing deep learning and optimization-based methods.

The rest of this study is organized as follows. Section II extensively examines the existing research related to hate speech detection. Section III provides an overview of the development and application of the proposed approach. Section IV discusses our experiments and the results of the proposed model. Finally, conclusions and potential future research directions are discussed in Section V.

II. Related Work

Research in NLP has progressed rapidly owing to the development of the social web. The social web has enabled researchers to use deep learning techniques along with NLP to interpret and understand social media-generated content, which has provided a large amount of data for researchers. Consequently, the social web has been established as an ever-changing environment for research. Social media is increasingly becoming multilingual; therefore, people from all around the world are able to communicate on different topics, such as current events and pop culture. However, while this linguistic diversity has provided many opportunities for communication, it has also increased the risk of encountering hate speech. Therefore, one of the biggest challenges in the management of online interactions is the ability to manage content across multiple languages.

Harmful and offensive forms of language, including hate speech, have been the focus of many studies. Hate speech can be defined as a form of communication that contains a prejudiced message that uses abusive language to target a person or people belonging to a particular group. The most prominent theme in hate speech research is the concept of targeting. This refers to the use of hate speech to specifically target identified groups of people (such as refugees) and communities.

Waseem et al. [10] identified four subcategories of offensive language based on the extent to which a post was directed toward a group of people (versus an individual), in addition to the level of specificity of the offensive language used. ElSherief et al. [11] focused on how users who posted hate messages are associated with their visibility on social media. The results showed that users with high levels of social media visibility also tended to be tolerant of hate speech. Finally, Salminen et al. [8] noted that hostile comments were common in online news discussion forums, especially those directed at news organizations and law enforcement. In general, researchers believe that online news discussions provide a suitable platform for the development of malicious behaviors[12].

A comprehensive study of the interactions between various social groups is essential for investigating the presence of hate speech on online platforms [13]. Various researchers have examined several facets of this subject matter, including the expressions of intolerance that are based upon social groups, how hateful content is strategically used as an effective persuasive device to disseminate offensive language, how the sharing of harmful content by social media users can increase social polarization, the spreading of hate within society, and the impact of social networks' structural characteristics. Given the profound influence of contextual and subjective factors, using interpretive techniques is a common approach to exploring nuances [14]-[17]. Burnap et al. [18] presented an innovative method for recognizing hate speech. Their method involves a hate-word-based feature selection process to select the characteristics required for the preferred embedding approach. Additionally, Basak et al. [4] developed a web application, "block shame," designed to identify and combat shaming online globally. This tool effectively silences and blocks spammers while offering a comprehensive definition of shaming encompassing six forms of abusive behavior: comparisons, judgmental comments, sarcasm, jokes, and whataboutery.

Kumar et al. [19] discussed a novel method for optimizing hyperparameters using optimization algorithms and LSTM neural networks. They have tuned five hyperparameters of the deep learning neural network layer to enhance the accuracy of the proposed model. Pré-trained models have proposed an idea to enhance performance and reduce computational time. A deep-learning strategy based on recurrent neural networks (RNNs) was introduced for smaller datasets [20]. The categorization of human opinions is a direct outcome of utilizing a sophisticated attention tool in combination with multiple learning tasks.

Sequeira et al. [21] applied deep learning models and language-embedding techniques to categorize messages linked to drug misuse, as detailed in a previous study [22]. Using deep neural networks applied to autoencoders (AE) for detecting hate speech has effectively managed complex data. The SentiDiff method was developed to identify samples of tweet vagueness by leveraging transfer learning in conjunction with a deep CNN approach.

Wang et al. [23] proposed a deep model incorporating a tree construct at the local level, which expressed significant accuracy in sentiment recognition. Additionally, Chen et al. [24] have developed a single classification system aimed at detecting rumors within online societies, recognizing the role of rumors in encouraging prejudice. The use of sentence-level sentiment classification has shown good results owing to its ability to effectively handle the amount of noise often found in social media data [25]. Additionally, the combination of different types of computational methods allows for improved accuracy and a better understanding of detected sentiments [26]. In the area of hate speech detection, Yang et al. [27] suggested a deep learning-based fusion method using a transformer encoder model and derived a language model to detect hate speech. Watanabe et al. [28] examined approaches for identifying the inciting language using feature-extraction methods. Shah et al. [29] employed a hybrid fuzzy rule structure approach within a logic framework to determine hate speech from insecure text.

Deep learning techniques have been used for many applications and are increasingly being applied in Natural Language Processing (NLP). Recent studies have demonstrated improved performance of Deep Learning techniques compared to traditional Machine Learning models [30]. Ensemble learning is a technique in which two or more models can be combined to achieve better results. One researcher examined ensemble learning for the classification of Hate Speech and Vulgar Content and achieved an accuracy of 87%. The data from this study indicated that there was still potential to optimize the hyperparameters of the meta-classifier(s), which may improve the overall model performance. Another researcher utilized CNNs to classify hate-related words/expressions from a text dataset [43].

An alternative strategy for enhancing the performance of this task would be to incorporate bidirectional long short-term memories Memory) that provide enhanced insight into the sequential nature of messages. Zhang et al. [33] utilized sophisticated Deep Learning approaches to achieve strong type differentiation of character-level classification. RNN-based architectures for recognizing abusive language through Bangla Speech Recognition have been employed in similar studies [34]. Additionally, Zamperi et al. [35] and Golbeck et al. [36], the use of Deep Learning strategies to identify abusive content on various social media platforms has proven effective, as demonstrated and supported by the results. These investigations incorporated learning algorithms, such as Bi-GRU, RNN, and CNN.

Some studies have introduced optimization techniques such as Genetic Algorithms (GA) and Particle Swarm Optimization (PSO) to improve model performance. However, these methods suffer from issues like slow convergence, getting trapped in local optima, and an inadequate balance between exploration and exploitation. As a result, they are not always able to achieve optimal performance in complex tasks like hate speech detection. Additionally, many existing works are limited to single datasets or platforms, which reduces their ability to generalize across diverse and dynamic social media environments. To address these limitations, this study proposes an SSA-DNN method to overcome these limitations. The Sparrow Search Algorithm (SSA) provides a more efficient and adaptive optimization strategy, enabling better exploration of the hyperparameter space and faster convergence toward optimal solutions. This reduces the dependency on human expertise and improves the overall robustness of the model. Moreover, by evaluating the model on a large multi-source dataset, the study ensures better generalization and real-world applicability.

III. Methodology

This section explored the Sparrow Search Algorithm, described the annotated dataset from diverse social media platforms, and analysis of Data-preprocessing, discuss the proposed hybrid approach.

A. Sparrow Search Algorithm

Optimization problems are observed easily in different fields of study, including engineering and finance, among others; thus, it is very crucial to arrive at the best solutions to the problems. The Sparrow Search Algorithm [33] is a comparatively new meta-heuristic optimization algorithm that is inspired by the foraging process of sparrows. As opposed to many conventional optimization techniques that may have difficulties with high levels of dimensions, the Sparrow Search Algorithm is intended to make use of the intelligence and versatility of sparrows to search the solution space successfully.

SSA is an inventive optimization technique based on the social aspects observed of sparrows during their movement in search of food. The foraging conduct of sparrows can be classified as a producer-scrounger, consisting of two main mechanisms: There is a dissemination mechanism that would work as an experimental mechanism, and an alert mechanism as well. In this model, producers, characterized by their notable fitness, are responsible for finding food sources, designating foraging locations, and providing guidance for scavenging routes. On the other hand, scroungers enhance their prevalent fitness by selectively following producers with numerous favorable fitness attributes, thus securing sustenance. Additionally, someone monitors the significance levels of producers, who are seen as obtaining resources via potentially unethical standards.

If a producer maintains an increased significance level, some scrounger populations may employ energetic food acquisition. When noticing a predator, sparrows swiftly emit a warning signal and relocate to a secure location. Sparrows in the population's prominent region display random movements toward their fellow sparrows. When the security threshold drops below a predefined warning value, producers must drive scroungers away from the dangerous location.

Considering the population of N sparrows, individual sparrows explore a search space with D dimensions. You can represent the spatial coordinates of individual sparrows as part of an NxD matrix. If you let $X(i,j)$ represent the location of the i^{th} sparrow in the j^{th} dimension, it is clear that i is 0 to N, and j is 0 to D. In this context, the sparrow population adheres to certain limitation conditions, where SSA selects 10%~15% of the producers. Eq. (1) is as follows:

$$X_{i,j}^{c+1} = \begin{cases} X_{i,j}^{c} * \exp\left(\frac{-i}{\alpha \cdot c_{max}}\right), A_V < TR \\ X_{i,j}^{c} + r.L \quad , A_V \geq TR \end{cases} \quad (1)$$

The variable "c" represents the current iteration count, and "α" stands for an arbitrary number uniformly distributed in the range of 0 to 1, "c_{max}" indicates the maximum number of iterations, and "r" is an arbitrary variable following a normal distribution. "L" is a 1 by D matrix consisting of elements set to 1.

The "A_V" expresses the alarm value, a single value between 0 and 1. Conversely, "TR" signifies the safety threshold, a single value between 0.5 and 1. If A_V is less than TR, there is no predator presence, and a comprehensive search can proceed. If A_V is equal to or greater than TR, it indicates the detection of a predator.

Consequently, the population needs to migrate to a safe place upon receipt of the alarm signal. Scroungers acquire food by trailing the producers, and the producer's location where abouts are adapted in Eq. (2).

$$X_{i,j}^{c+1} = \begin{cases} Q * \exp\left(\frac{X_{worst}^{c} - X_{i,j}^{c}}{i^2}\right), i > \frac{n}{2} \\ X_p^{c+1} + \left|X_{i,j}^{c} - X_p^{c+1}\right|.A^{+}.L \quad , i \leq \frac{n}{2} \end{cases} \quad (2)$$

The equation, as previously mentioned, involves multiple variables. Here, X_p represents the best place for the producer, and X_{worst}corresponds to the worst global status. The matrix denoted as A consists of randomly induced elements, each being either 1 or -1, and follows a specific equation. For instance, when the value of 'i' surpasses half of 'n,' it implies that the *i*^th^ individual scrounging for food is hungry, displaying a suboptimal physical situation and reduced vitality levels. Consequently, this individual must seek nourishment in different areas. In this context, the individual scrounging for food follows the producers in a considerably beneficial position. The matrix indicated is determined using the formula A^{+} is $A^T(AA^T)^{-1}$. Updating monitors in Eq. (3) as follows:

$$X_{i,j}^{c+1} = \begin{cases} X_{best}^{c} + \beta.\left|X_{i,j}^{c} - X_{best}^{c}\right|, \quad f_i > f_g \\ X_{i,j}^{c} + K.\left(\frac{\left|X_{i,j}^{c} - X_{worst}^{c}\right|}{(f_i - f_w) + \epsilon}\right), f_i = f_g \end{cases} \quad (3)$$

In this scenario, X_{best} represents the optimal location within the entire area. The parameters β and K order the step size, arbitrarily induced from a normal distribution between -1 and 1. The variable f_i describes the current fitness level of a sparrow. Additionally, f_r describes the sparrow's current fitness level, whereas the best and optimal positions across the total location are implied as $f_g$and f_w, respectively. To avoid division by zero, the constant ε is added. If, however, $f_i > f_g$, it implies that the sparrow is arranged at the population's periphery and thus vulnerable to predators. The sparrow must seek safety by joining different group members in such instances.

According to the literature review, SSA has been developed to overcome the existing meta-heuristic algorithms, which have no prior information about the solution space topography. The algorithm adds a method of visualizing the characteristics of the solution space by sets of solutions found, as one draws maps to improve the probability of discovering the treasure. This way, the algorithm can make better decisions during the search and thus solve the significant optimization problems efficiently.

The idea of the Sparrow Search Algorithm is based on the actual foraging behavior of sparrows as socially foraging birds that tend to use both individual and social approaches. Sparrows are very intelligent birds that are capable of adjusting to new surroundings and collaborating in order to gain the best results in terms of feeding. The functions of the operators of the algorithm – the selection of the search space, further search within a deviated search space, further search within a converged search space, swooping to capture prey – are modeled on these natural processes. The Sparrow Search Algorithm has been evaluated and benchmarked with other famous meta-heuristic algorithms like Particle Swarm Optimization, Genetic Algorithms, and Grey Wolf Optimizer for different optimization problems, including mathematical benchmark functions as well as some engineering application problems [19, 31-32]. These results have clearly proved that the Sparrow Search Algorithm has the ability to outperform or at least be at par with these recognized algorithms, thus explaining its accuracy and applicability.

Another major strength of the Sparrow Search Algorithm is its ability to adopt the exploration/exploitation trade-off during the search phase. The features of how the algorithm chooses the search space and how it adapts within a strayed search field and within a refined search field explain why this algorithm allows it to search the solution space and find the global optimum. Moreover, the implementation of parallel algorithms is also possible with the help of the developed algorithm, which also minimizes the computing time to achieve better results in solving various optimization problems.

B. Description of Annotated Dataset

In this study, we employed ten pre-existing annotated datasets [10], [34]-[42]. Five of these datasets originated from Twitter, two from YouTube/Facebook, one from Gab, one from Wikipedia, and the remaining from Reddit. Table I provides an overview of these datasets collected from diverse social media. These annotated datasets consisted of English-language communication texts and were categorized as either hate speech or non-hate. All datasets from different social media networks were divided into a unified dataset comprising 432053 items. Table II shows the text in this consolidated dataset, indicating whether they are classified as hate speech or non-hate speech.

C. Data Preprocessing

To gain an understanding of how to improve the proposed methodology's performance and enable its operation at the highest level possible, the data were systematically analyzed, categorized, and refined. Data analysis was used to develop generalizable characteristics from the available CSV dataset. During the first phase of pre-processing for this dataset, several undesired items were eliminated, such as emoticons, non-alphanumeric symbols ($, %, >), URL addresses, and 'noise' such as ;, &abc!, |, and :. Next, punctuation symbols were eliminated, and each hashtag (i.e., #BanAfrica) was converted to regular text by replacing the # symbol with ordinary characters. Following the normalization of the clean text data using stemming, lemmatization, and conversion to lower case, the data were tokenized, resulting in the identification of 89,419 unique tokens across the social media platforms. Examples of some pre-processing steps are listed in Table III.

TABLE. I. DESCRIPTION OF ANNOTATED DATASETS

S. No	Dataset	Data Source	Year	ML Approach	Dataset
1	Waseem and Hovy	Twitter [10]	2016	LR	16910
2	JIGSAW	Wikipedia [34]	2017	Corpus	250677
3	Golbeck	Twitter [36]	2017	Corpus	20360
4	Davidson	Twitter [37]	2017	LR, SVM, DT, NB	24783
5	Zampieri	Twitter [35]	2019	CNN	13240
6	Chung	Facebook [38]	2019	Corpus	20186
7	Hopf	YouTube [39]	2020	SVM	3222
8	HateXplain	Twitter, Gab [40]	2020	Corpus	20148
9	Derek	Reddit [41]	2020	Reddit Corpus	40000
10	Kennedy	Gab [42]	2022	Gab corpus	22527
Total					**432053**

TABLE. II. DESCRIPTION OF VARIOUS SOCIAL MEDIA INSTANCES

Class	Label	Instances
Hate Speech	0	161623
Neither	1	270430

TABLE. III. STEPS OF DATA-PREPROCESSING

Text Pre-processing	Textual Data
“My Grandma used to call me a porch Monkey all the time... Come to think of it she did refer to a broken bottle as a nigger knife” &#128563;	
Filtering	“My Grandma used to call me a porch Monkey all the time Come to think of it she did refer to a broken bottle as a nigger knife” #128563;
Remove Stop-Words and Hashtags	Grandma call porch Monkey time think broken bottle nigger knife
Lowercase	grandma call porch monkey time think broken bottle nigger knife
Stemming and Lemmatization	grandma call porch monkey time think broke bottle nigger knife
Tokenizer	[‘ruin’, ‘life’, ‘throw’, ‘smoke’, ‘weed’, ‘test’, ‘score’, ‘better’, ‘discard’, ‘life’] [‘grandma’, ‘call’, ‘porch’, ‘monkey’, ‘time’, ‘think’, ‘broke’, ‘bottle’, ‘nigger’, ‘knife’]

D. Proposed Approach

The SSA is an efficient optimization algorithm that is easily controlled because it uses few control parameters. The SSA uses a random search strategy to define the initial position of the sparrows. This can lead to longer convergence times and less accurate solutions if the randomly chosen initial positions are significantly away from their respective optimal positions. However, SSA has proven to be effective in solving multivariable complex optimization problems; that is, SSA has demonstrated the ability to quickly find the optimal solution and provide reliable results. Research has been conducted on using SSA with other optimization problems in various fields of engineering, which provides a basis for the current investigation. Unlike previous studies [34], the objective of this study was to improve the adaptability of the DNN-based learning model to human influence through hyperparameter tuning, as well as improve the model's accuracy in making predictions. The key parameters of the DNN, that is, the learning rate, number of LSTMs used in the network, number of neurons in the dense layer, and number of training epochs, were systematically searched for optimal values. The SSA was used to optimize these hyperparameters to ensure that consistent features were extracted and a stable network architecture was created.

The ability of SSA to effectively determine the optimal parameter weights for deep learning algorithms improves their interpretability. The SSA-DNN model, depicted in Fig. 1, is proposed to train optimized deep neural networks. This architecture consists of four separate layers: an input, an LSTM [1], a hidden and an output. The SSA focuses on optimizing parameters such as the learning rate, number of neurons in two hidden layers, and the corresponding number of optimized epochs. Once the parameter scale is set, the positional facts of the population, as well as the parameters corresponding to them, are set randomly. This study aims to examine the fitness function employed in SSA as well as the loss method of the deep neural network, which belongs to the framework of the learning model. In this approach, each dataset was randomly divided into two parts: 80% for training and 20% for testing purposes. All experiments in this study were conducted using Python 3.8 and TensorFlow 2.6.0, running on a Windows 10 system equipped with an NVIDIA Quadro K2220 GPU (32 GB) and a dual Intel Xeon E5-1650 CPU at 3.50 GHz, supported by 64 GB of RAM.

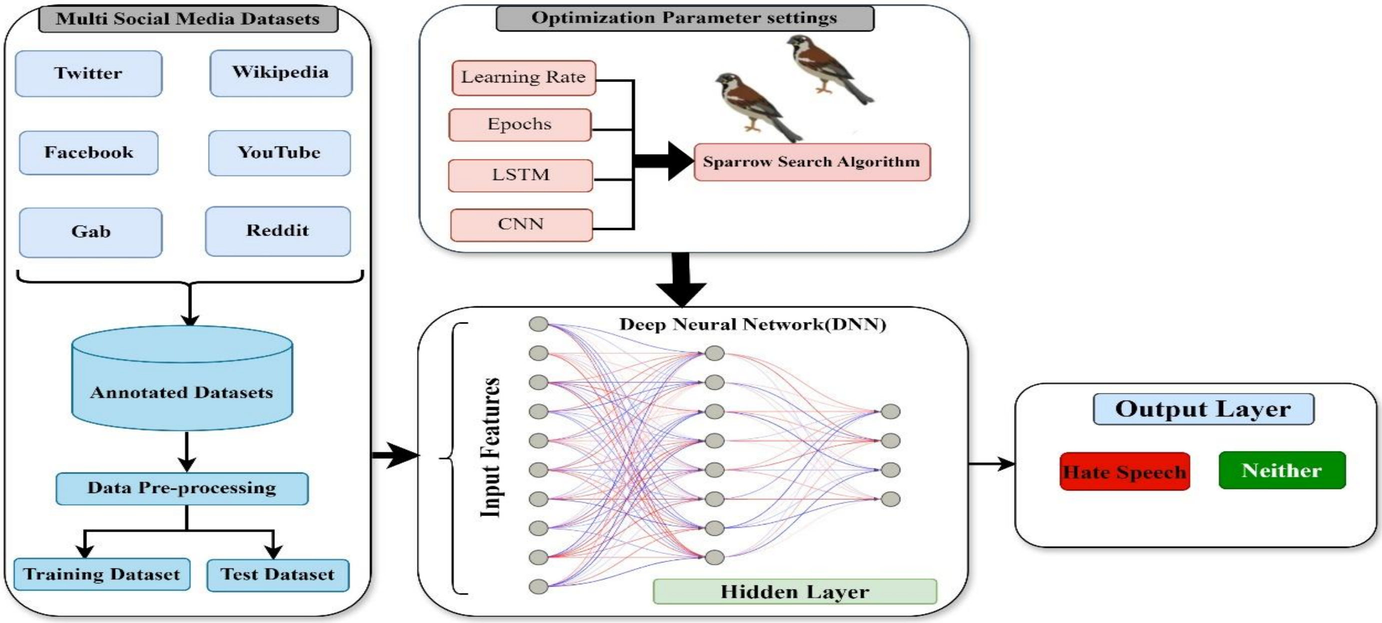


Fig. 1. Proposed hybrid architecture.

IV. Results and Discussion

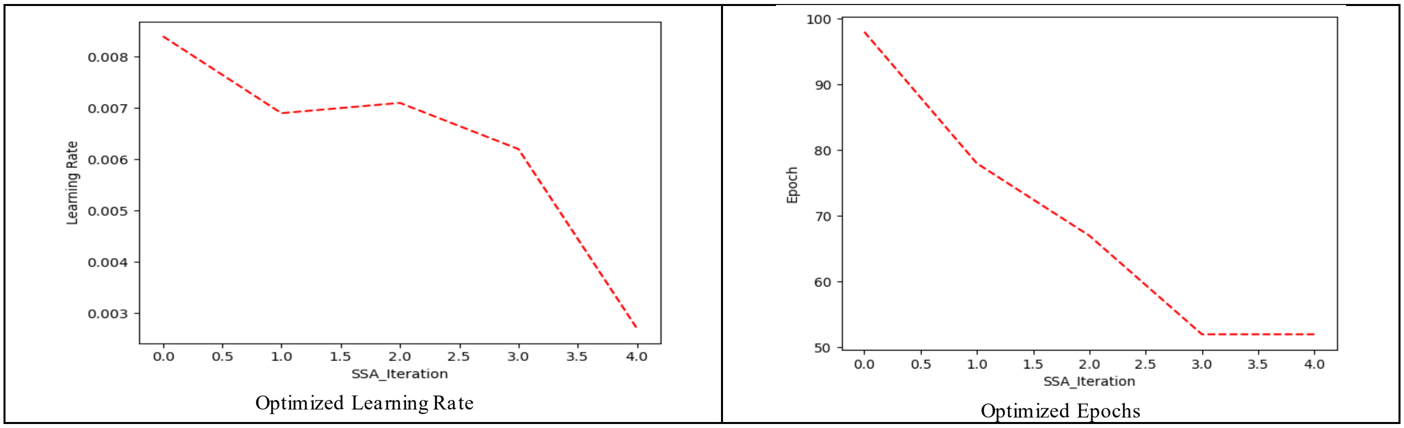
This section explored the optimization of the hyperparameters of our SSA-DNN model, evaluation of our SSA-DNN model across various datasets, and compared it with state-of-the-art methods.

A. Optimizing Hyperparameters on Proposed Model

The hyperparameters of the proposed model that is responsive to optimization using the Sparrow Search Algorithm consist of the following:

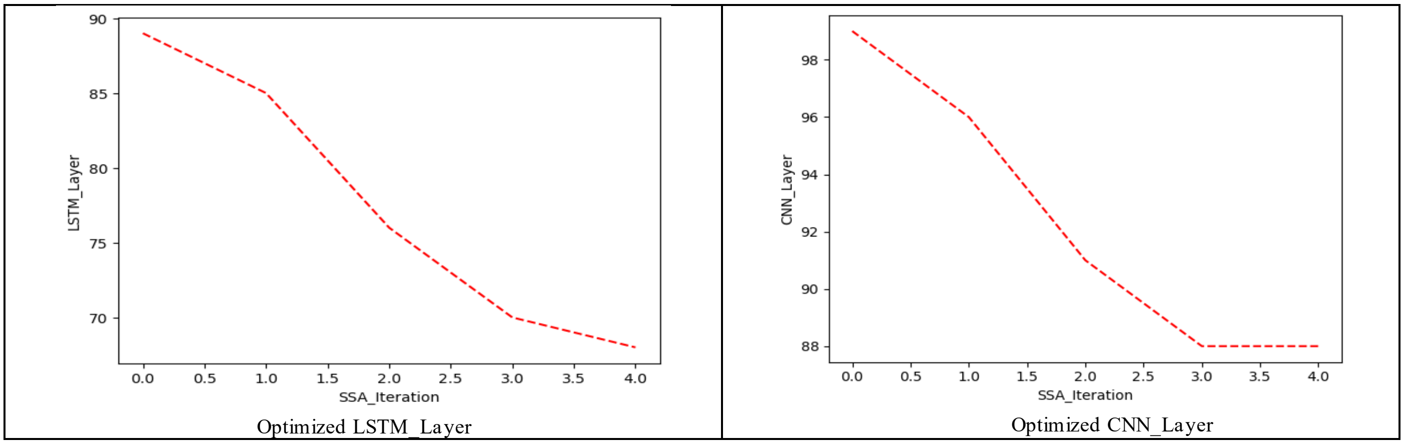
- neurons in the LSTM layer
- values of the learning rate
- neurons in the CNN layer
- epochs

A concise overview of these hyperparameter values is provided in Fig. 2. The optimal hyperparameters for the Deep neural network were determined based on the information in Table IV. These optimized parameter settings include a learning rate of 0.0027, an epoch count of 52, 68 neurons in the LSTM layer, and 88 neurons in the CNN layer. The proposed model for identifying hate speech messages was developed using the most effective parameter weights acquired through optimization using a sparrow search algorithm.



Optimized Learning Rate

Optimized Epochs



Optimized LSTM_Layer

Optimized CNN_Layer

Fig. 2. Fine-tuning hyperparameters of the proposed hybrid model.

TABLE. IV. FINE-TUNING OF THE PROPOSED MODEL

Parameters	Scale	Best value
Learning_Rate	0.0001 – 0.01	0.0027
Epoch	0-80	52
DNN_LSTM	0-128	68
DNN_CNN	0-128	88

B. Results

Table V elaborates the classification outcomes arising from the proposed model during the iterations. The results of the above iterations show that the proposed model effectively classifies messages in various epochs. Thus, in iteration 1, the accuracy of the proposed model was 95.6. After iteration 3, the accuracy level of the proposed model was 97.5. As can be seen from Iteration 4 of the proposed model, the prediction results accurately reached 97.9. Finally, after the fifth iteration, a model with the proposed solution was created with an accuracy of 97.9.

TABLE. V. RESULTS OF THE PROPOSED MODEL

Iteration	Accuracy	P	R	F1_Score
1	95.6	93.8	94.4	94.1
2	96.9	95.5	96.4	95.9
3	97.5	96.1	97.2	96.6
4	97.9	97.0	97.8	97.4
5	**97.9**	**97.0**	**97.8**	**97.4**

Table VI presents a comparative study of the proposed model and the current state-of-the-art model. Initially, the performance metrics of techniques designed by [35], [39] utilizing machine learning approaches exhibited relatively lower accuracy levels, with values of 93.1 and 93.0, respectively. Next, the approaches presented by [41],[44], incorporating deep learning approaches, were performed with remarkable precisions of 91.2 and 87.0. However, we also provide the result of applying the same GA and PSO algorithm to optimize the suggested model, with less accuracy, 74.5 and 78.6, as compared to the SSA-DNN model. Ultimately, the proposed model attained its extreme accuracy of 97.9. These outcomes emphasize the outstanding implementation of the proposed model compared to other deep learning approaches, resulting in enhanced performance.

TABLE. VI. PROPOSED MODEL VS STATE OF ART METHODS

Authors	Methods	P	R	Accuracy
Vashistha et al. [44]	LR, TF-IDF	91.2	91.2	91.0
Kurrek et al. [41]	CNN, LSTM	90.1	92.0	87.0
Hopf et al. [39]	SSA	-	-	93.1
Kumar et al. [19]	LSTM, SSA	93.5	93.6	93.6
Zampieri et al. [35]	LSTM, GBDT	93.0	93.0	93.0
Genetic Algorithm (GA)	LSTM-CNN	74.5	74.6	74.5
Particle Swarm Optimization (PSO)	LSTM	78.2	79.0	78.6
Proposed Model	**SSA, DNN**	**97.0**	**97.8**	**97.9**

V. CONCLUSION

This study introduced a hybrid SSA-DNN framework for hate speech detection across multiple social media platforms, with a primary focus on automated and efficient hyperparameter optimization. By integrating the Sparrow Search Algorithm (SSA) with deep neural networks, the proposed approach overcomes key limitations of traditional tuning methods, which often depend on manual expertise and exhaustive search strategies. The SSA mechanism enables intelligent exploration of the hyperparameter space, leading to the optimal selection of critical parameters such as learning rate, number of neurons, and training epochs. This not only minimizes human intervention but also enhances the robustness and generalization capability of the model. The effectiveness of the proposed framework is validated through extensive experiments conducted on a large-scale, multi-source dataset comprising diverse social media platforms. The results demonstrate that the SSA-DNN model achieves a high classification accuracy of **97.9%,** outperforming several existing machine learning and deep learning approaches. Additionally, the model shows improved precision, recall, and F1-score, indicating its capability to reliably distinguish between hate and non-hate content. Beyond performance improvements, the SSA-based optimization process contributes to better interpretability by offering insights into parameter selection and model behavior. This makes the proposed approach not only effective but also more transparent compared to conventional deep learning models.

Despite these promising outcomes, there are several avenues for future research. One important direction is the enhancement of computational efficiency by reducing training time and exploring lightweight neural architectures suitable for real-time applications. As social media platforms generate massive volumes of data continuously, scalable and efficient models are essential for practical deployment. Another significant extension involves the incorporation of multimodal data**,** including images, videos, and audio, alongside textual information. Hate speech is often expressed through memes, visual symbols, or spoken language, which cannot be fully captured by text-only models. A multimodal SSA-DNN framework could therefore improve detection accuracy by capturing such nuanced and context-dependent forms of harmful content. Furthermore, integrating the proposed framework with transfer learning and pre-trained language models**,** such as transformer-based architectures, could enhance performance across low-resource languages and domain-specific datasets. This would increase the adaptability and applicability of the model in diverse linguistic and cultural settings. Future work may also explore the integration of explainable AI techniques to provide deeper insights into model decisions, thereby ensuring fairness, accountability, and bias mitigation in automated hate speech detection systems.

REFERENCES

[1] Hochreiter S, Schmidhuber J. LSTM can solve hard long time lag problems. Advances in neural information processing systems. 1996.

[2] Ousidhoum, N. et al." Multilingual and multi-aspect hate speech analysis". arXiv preprint arXiv:1908.11049,2019.

[3] Kumar, A et al. "Optimized Deep Neural Networks Using Sparrow Search Algorithms for Hate Speech Detection." International Journal of Computing and Digital Systems 15.1 (2024): 1-9.

[4] R. Basak, S. Sural, N. Ganguly, and S. K. Ghosh, "Online public shaming on twitter: Detection, analysis, and mitigation," IEEE Transactions on Computational Social Systems, vol. 6, no. 2, pp. 208–220, 2019

[5] Waseem Z. Are you a racist or am i seeing things? Annotator influence on hate speech detection on twitter. In:Proceedings of the first workshop on NLP and computational social science; 2016. P. 138–142

[6] Badjatiya P, et al. Deep learning for hate speech detection in tweets. In: Proceedings of the 26th international conference on world wide web companion, Geneva; 2017. P. 759–60.

[7] Mondal M, et al. A measurement study of hate speech in social media. In: Proceedings of the 28th ACM conference on hypertext and social media, New York; 2017. P. 85–94.

[8] Salminen, et al. Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In: Proceedings of the international AAAI conference on web and social media (ICWSM 2018), San Francisco; 2018.

[9] Herring S et al (2002) Searching for safety online: managing" trolling" in a feminist forum. Inf Soc 18(5):371–384

[10] Waseem Z, et al. Understanding abuse: a typology of abusive language detection subtasks. arXiv :1705.09899 [cs].2017.

[11] ElSherief M, et al. Peer to peer hate: hate speech instigators and their targets. In: Proceedings of the twelfth international AAAI conference on web and social media, Palo Alto; 2018.

[12] Malmasi, S.; Zampieri, M. Challenges in discriminating profanity from hate speech. J. Exp. Theor. Artif. Intell. 2018, 30, 187–202.

[13] Al-Ajlan, M.A.; Ykhlef, M. Optimized Twitter Cyberbullying Detection based on Deep Learning. In Proceedings of the 2018 21st Saudi Computer Society National Computer Conference (NCC), Riyadh, Saudi Arabia, 25–26 April 2018; pp.1–5.

[14] Ahmed, M.T.; Rahman, M.; Nur, S.; Islam, A.; Das, D. Deployment of Machine Learning and Deep Learning Algorithms in Detecting Cyberbullying in Bangla and Romanized Bangla text: A Comparative Study. In Proceedings of the 2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), Bhilai, India, 19–20 February 2021.

[15] Dadvar, M.; Eckert, K. Cyberbullying detection in social networks using deep learning-based models. In International Conference on Big Data Analytics and Knowledge Discovery; Springer: Cham, Switzerland, 2020.

[16] Luo, X. Efficient English text classification using selected machine learning techniques. Alex. Eng. J. 2021, 60, 3401–3409.

[17] Khan, U.; Khan, S.; Rizwan, A.; Atteia, G.; Jamjoom, M.M.; Samee, N.A. Aggression Detection in social media from Textual Data Using Deep Learning Models. Appl. Sci. 2022, 12, 5083.

[18] W. Alorainy, P. Burnap, H. Liu, and M. L. Williams, "the enemy among us": Detecting cyber hate speech with threats-based other language embeddings," ACM Transactions on the Web, vol. 13, no. 3, pp. 1–26, Jul. 2019.

[19] Kumar, S. et al. NDDSM: Novel Deep Decision-Support Model for Hate Speech Detection. SN COMPUT. SCI. 5, 67 (2024).

[20] L. Gui, L. Jia, J. Zhou, R. Xu, and Y. He, "multi-task learning with mutual learning for joint sentiment classification and topic detection," IEEE Transactions on Knowledge and Data Engineering, pp. 1–13, 2020.

[21] R. Sequeira, A. Gayen, N. Ganguly, S. K. Dandapat, and J. Chandra, "A large-scale study of the twitter follower network to characterize the spread of prescription drug abuse tweets," IEEE Transactions on Computational Social Systems, vol. 6, no. 6, pp. 1232–1244, 2019.

[22] R. Zhao and K. Mao, "Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder," IEEE Transactions on Affective Computing, vol. 8, no. 3, pp. 328–339, 2017.

[23] J. Wang, L. Yu, K. R. Lai, and X. Zhang, "Tree-structured regional cnn-lstm model for dimensional sentiment analysis," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 581–591, 2020.

[24] A. Ebrahimi Fard, M. Mohammadi, Y. Chen, and B. Van de Walle, "Computational rumor detection without non-rumor: A one-class classification approach," IEEE Transactions on Computational Social Systems, vol. 6, no. 5, pp. 830–846, 2019.

[25] L. G. Singh, A. Anil, and S. R. Singh, "She: Sentiment hashtag embedding through multitask learning," IEEE Transactions on Computational Social Systems, vol. 7, no. 2, pp. 417–424, 2020.

[26] K. Rudra, A. Sharma, N. Ganguly, and S. Ghosh, "Characterizing and countering communal microblogs during disaster events," IEEE Transactions on Computational Social Systems, vol. 5, no. 2, pp. 403–417, 2018.

[27] Y. Zhou, Y. Yang, H. Liu, X. Liu, and N. Savage, "Deep learning-based fusion approach for hate speech detection," IEEE Access, vol. 8, pp. 128 923–128 929,2020.

[28] H. Watanabe, M. Bouazizi, and T. Ohtsuki, “Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection,” IEEE Access, vol. 6, pp. 13 825–13 835, 2018.

[29] P. Shah and A. Patel, "A Comprehensive Study and Detailed Review on Hate Speech Classification: A Systematic Analysis," 2021 International Conference on Advances in Computing, Communication, and Control (ICAC3), Mumbai, India, 2021, pp. 1-6, doi: 10.1109/ICAC353642.2021.9697110.

[30] Kumar A, Kumar S, Passi K, Mahanti A. A hybrid deep BiLSTM-CNN for hate speech detection in multi-social media. ACM Transactions on Asian and Low-Resource Language Information Processing, 2024.

[31] Liu F, Qin P, You J, Fu Y. Sparrow Search Algorithm-Optimized Long Short-Term Memory Model for Stock Trend Prediction. Computational Intelligence Neuroscience. 2022 Aug 12; 2022:3680419. Doi: 10.1155/2022/3680419. PMID: 35990139; PMCID: PMC9391098.

[32] G. I. Rajathi, R. R. Kumar, D. Ravikumar, T. Joel, S. Kadry et al., “Brain tumor diagnosis using sparrow search algorithm based deep learning model,” Computer Systems Science and Engineering (CSSE), vol. 44, no. 2, pp. 1793– 1806, 2023.

[33] C. L. Zhang and S. F. Ding, “A stochastic configuration network based on chaotic sparrow search algorithm,” Knowledge-Based Systems, vol. 220, Article ID 106924, 2021.

[34] cjadams, Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, nithum, Will Cukierski. (2017). Toxic Comment Classification Challenge. Kaggle. https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge.

[35] Zampieri, M. et al.,”Predicting the type and target of offensive posts in social media, in: NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies – Proceedings of the Conference. Association for Computational Linguistics (ACL), pp. 1415–1420,2019

[36] Golbeck, J. et al., “A large, labeled corpus for online harassment research”. In Proceedings of the 2017 ACM on web science conference, pp.229-233,2017.

[37] Davidson, T. et al., “Automated hate speech detection and the problem of offensive language”, in Proceedings of the 11th International Conference on Web and social media, ICWSM 2017. AAAI Press, pp. 512–515,2017.

[38] Chung, Y.L. et al., “CONAN-Counter Narratives through Nichesourcing: A Multilingual Dataset of Responses to Fight Online Hate Speech”. arXiv preprint arXiv:1910.03270,2019.

[39] Hopf, M. et al.,” Developing an online hate classifier for multiple social media platforms. Human-centric Computing and Information Sciences,2020.

[40] Mathew, B.; Saha, P.; Yimam, S.M.; Biemann, C.; Goyal, P.; Mukherjee, A. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. arXiv 2020, arXiv:2012.10289.

[41] Kurrek, J. et al., “Towards a comprehensive taxonomy and large-scale annotated corpus for online slur usage”. In Proceedings of the Fourth Workshop on Online Abuse and Harms, pp. 138-149,2020.

[42] Kennedy, B. et al., “Introducing the Gab Hate Corpus: defining and applying hate-based rhetoric to social media posts at scale. Language Resources and Evaluation, 56(1), pp.79-108,2022.

[43] Gambäck, Björn, and Utpal Kumar Sikdar. "Using convolutional neural networks to classify hate-speech." In Proceedings of the first workshop on abusive language online, pp. 85-90. 2017.

[44] Vashistha, Neeraj, and Arkaitz Zubiaga. "Online multilingual hate speech detection: experimenting with Hindi and English social media." Information 12.1, 2020.