Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.047840

ARTICLE

An Adaptive Hate Speech Detection Approach Using Neutrosophic Neural Networks for Social Media Forensics

Yasmine M. Ibrahim[1,2], Reem Essameldin[3] and Saad M. Darwish[1,*]

[1]Department of Information Technology, Institute of Graduate Studies and Research, Alexandria University, Alexandria, 21526, Egypt

[2]Faculty of Computers and Information Technology, Egyptian E-Learning University (EELU), Giza, 12611, Egypt

[3]Faculty of Computers and Data Science, Alexandria University, Alexandria, 21554, Egypt

*Corresponding Author: Saad M. Darwish. Email: saad.darwish@alexu.edu.eg

Received: 20 November 2023 Accepted: 04 January 2024 Published: 25 April 2024

ABSTRACT

Detecting hate speech automatically in social media forensics has emerged as a highly challenging task due to the complex nature of language used in such platforms. Currently, several methods exist for classifying hate speech, but they still suffer from ambiguity when differentiating between hateful and offensive content and they also lack accuracy. The work suggested in this paper uses a combination of the Whale Optimization Algorithm (WOA) and Particle Swarm Optimization (PSO) to adjust the weights of two Multi-Layer Perceptron (MLPs) for neutrosophic sets classification. During the training process of the MLP, the WOA is employed to explore and determine the optimal set of weights. The PSO algorithm adjusts the weights to optimize the performance of the MLP as fine-tuning. Additionally, in this approach, two separate MLP models are employed. One MLP is dedicated to predicting degrees of truth membership, while the other MLP focuses on predicting degrees of false membership. The difference between these memberships quantifies uncertainty, indicating the degree of indeterminacy in predictions. The experimental results indicate the superior performance of our model compared to previous work when evaluated on the Davidson dataset.

KEYWORDS

Hate speech detection; whale optimization; neutrosophic sets; social media forensics

1 Introduction

Nowadays, Online Social Networks (OSNs) are becoming more popular among individuals all over the world whether they are technological or not. One of the main reasons for OSNs widespread usage is the accessibility of the internet. Individuals use OSN for a variety of reasons, including entertainment, interaction, joy, fame, marketing, and commerce [1]. OSN allowed a fertile medium to spread hateful, harmful, and aggressive content. Despite differences in hate speech laws in different countries, it commonly includes communications of hatred against individuals or groups based on a group feature such as religion, color, gender, or disability. The proliferation of hate speech on social media platforms has exacerbated the challenges confronted by investigators [2].

Social media forensics is an approach to gathering, analyzing, and investigating digital information collected through different social media platforms to find evidence for court or criminal investigations. In digital forensics, social media evidence is a fresh field [3]. Throughout social media platforms criminal activities and evidence can be detected, and the data collected from applications and platforms can be used in criminal investigation. Social media forensics analysis steps include the following [4–7]: (a) Collection of social media content (Evidence Collection): This can be accomplished through a variety of methods including web scraping, Application Programming Interfaces (APIs), and specialist tools developed for social media data collection. Text, photos, videos, metadata, and other information important to the analysis can all be included in the evidence [6]. (b) Data preprocessing: This might include cleaning up the data from duplicates, unnecessary data, and additional noise. (c) Data Analysis: The next stage is to analyze the data using many techniques and tools. These techniques encompass a variety of approaches, including Natural Language Processing (NLP) [8] and text classification for analyzing data to find patterns, trends, and other insights that can inform the objectives of the analysis. Like the classification of text whether it contains hate speech, cyberbullying, and offensive speech or not. (d) Data Interpretation of Digital Evidence: This involves making sense of the patterns, trends, and other insights that were uncovered during the analysis. (e) Reporting: This may involve creating a report, a dashboard, or other forms of visual representation of the data and insights.

Hate speech detection for social media forensics can contribute to mitigating the presence of harmful content on social media platforms, preventing violence, and fostering social harmony [9]. The process of detecting hate speech involves discerning the language employed in social media posts that target individuals or groups based on characteristics such as race, religion, ethnicity, gender, sexual orientation, or other attributes [10]. Detecting hate speech is a challenging task as ambiguity in any language varies greatly depending on the speaker and the audience [11]. Ambiguity refers to a text that has multiple meanings or interpretations, making it difficult to determine the intended meaning without additional context or clarification [12]. Text classification [13] is one of the most important and active research areas in the social media analysis domain. It is a machine-learning task that involves automatically assigning text to a predefined category. This is a valuable tool in the social media domain, where it can be used to classify social media posts into different categories, such as "hate speech", "cyberbullying", or "offensive speech".

There are three main approaches to hate speech detection [14]: keyword-based, machine learning-based, and hybrid approaches. It relies on the identification of specific words or phrases that are known to be associated with hate speech. The keyword-based approach is relatively simple to implement, but it can be easily circumvented by users who use euphemisms or slang terms. Furthermore, the machine learning approach uses statistical models to learn the patterns of language that are associated with hate speech. It is more robust to evasion than the keyword-based approach. However, obtaining a substantial amount of labeled data for training and relying on human experts to label social media posts as hate speech or not hate speech are necessary but challenging aspects. A hybrid approach combining keyword-based and machine learning-based methods has been proposed.

Within the second category, the Multi-Layer Perceptron (MLP) [15] is broadly used. MLP classification is an effective machine learning technique used for the categorization of hate speech in textual data. MLP classifiers are composed of multiple layers of artificial neurons, which are interconnected in a specific way. Within an MLP classification model, the neurons in each layer are responsible for learning distinct features present in the input text. The output of the final layer is then utilized to classify the input text into various categories, such as hate speech or non-hate speech. MLPs offer advantages in handling complex nonlinear relationships between features, making

them suitable for text classification tasks. MLPs demonstrate scalability by accommodating large volumes of data. This scalability is achieved by increasing the number of layers and neurons in the network, enabling MLPs to effectively learn intricate relationships between input features and output classes. This capability is particularly beneficial in text classification tasks [15]. Recently, the Whale Optimization Algorithm (WOA) [16] is an effective algorithm for optimizing the training of MLPs. The WOA algorithm is a population-based algorithm, which means that it uses a group of search agents to search for the optimal solution. The WOA algorithm draws inspiration from the natural behavior of humpback whales, making it a bio-inspired optimization algorithm. It has shown effectiveness in solving diverse optimization problems, including function optimization, machine learning, and scheduling problems. Notably, WOA exhibits several advantages over other optimization algorithms, such as rapid and efficient exploration of large search spaces, faster convergence to optimal solutions, and reduced sensitivity to data noise [16].

The subjective nature of language in speech can make it difficult to identify and classify whether it contains hate speech or not. This is because the meaning of a text can depend on a variety of factors, including the context in which it is used, the intent of the speaker or writer, and the cultural background of the audience [17]. Despite these challenges, it is important to continue to develop tools and techniques for automatically identifying and classifying hate speech. Hate speech can harm individuals and society, and it is important to be able to identify and remove it from online platforms [18].

Neutrosophic Logic (NL) [19] emerges as a promising and innovative tool for hate speech detection, showcasing unique features that set it apart from traditional fuzzy tools. NL's distinct capability lies in its nuanced representation of uncertainty, surpassing conventional fuzzy tools. By employing a three-valued membership function, NL captures degrees of truth, indeterminacy, and falsity, offering a sophisticated understanding of linguistic nuances in hate speech instances. Comparatively, Neutrosophic Logic employs a three-valued representation, capturing degrees of truth, falsity, and indeterminacy. In contrast, Traditional Machine Learning is typically limited to truth and falsity, while Deep Learning relies on probabilities. Fuzzy Logic [20,21] represents uncertainty through degrees of membership and non-membership. The nuanced approach of NL, with its explicit representation of indeterminacy and membership functions, positions it as a tool capable of addressing the complexities of hate speech detection more effectively than traditional fuzzy tools, and machine learning. This comparative analysis highlights the unique strengths of Neutrosophic Logic in the context of hate speech classification.

1.1 Problem Statement

Hate speech poses a growing challenge on social media platforms, as it can quickly propagate and inflict harm upon individuals and communities. The absence of robust mechanisms for monitoring and reporting hate speech on these platforms has contributed to a rise in online incidents involving hate speech. In the realm of social media forensics, the identification and analysis of hate speech plays a pivotal role in comprehending and mitigating its dissemination. Manual monitoring and reporting of hate speech on social media platforms are resource-intensive, time-consuming, and often ineffective. Consequently, there is a pressing need to develop automated systems for hate speech detection that can accurately and efficiently identify hate speech in real-time, thereby aiding social media forensics efforts.

1.2 Motivation

Detecting hate speech automatically on social media has emerged as an exceptionally challenging task due to the complex nature of language, these challenges include unwanted content, such as meaningless texts that can hamper the performance of detection algorithms. Differentiating between contaminated content and real-world events is a key obstacle. Scalable and effective metadata architectures are essential for real-time hate speech detection. Short messages and grammatical errors further complicate traditional text analysis techniques. The brevity and informal nature of text often result in ambiguous or unclear expressions, making it challenging to accurately interpret the intended meaning. Identifying hate speech becomes even more complex when faced with uncertain or contextually vague content. Currently, several methods exist for hate speech classification like single machine learning methods and hybrid machine learning methods, but they still suffer from ambiguity when differentiating between hateful and offensive content and they also lack accuracy.

1.3 Contribution and Methodology

This paper offers a hate speech classification model based on neutrosophic neural networks. The model combines WOA and PSO [22] to adjust weights during the training of two MLPs for neutrosophic sets classification. The WOA explores optimal weights, while the PSO fine-tunes MLP performance. Two MLPs are employed; one predicts degrees of truth membership, the other false membership. The model estimates indeterminacy membership by calculating the difference between truth and false memberships, generating interval neutrosophic sets. Semantically, the model predicts the meaning of input patterns by associating features with truth, false, and indeterminacy memberships for each class. Trained on labeled input patterns, it handles ambiguity inherent in neutrosophic sets, accommodating uncertainty in membership values and the possibility of belonging to multiple classes simultaneously.

The paper's structure is as follows: Section 2 provides an overview of recent related work. Section 3 outlines the proposed model for hate speech detection. Section 4 presents the results and discussions. Finally, Section 5 presents the conclusions drawn from the study.

2 Related Works

Numerous studies have been conducted by researchers focusing on techniques for detecting hate speech on social media. This section presents a brief review of recent studies that employed various methods, such as Neural Network Models, Fuzzy, ensemble, and Hybrid models [14,18]. In [23], a proposed Twitter hate speech classification model used a rule-based system and modified Jaccard similarity for real-time tweet clustering, reducing misclassification. Results showed effectiveness in topic detection and categorization, demonstrating robustness to various challenges. However, the model displayed sensitivity to larger training datasets, escalating complexity. Moreover, in [24], authors suggested a hybrid approach combining lexicon-based and machine learning methods for hate speech prediction, integrating sentiment analysis. The limitations of this model include the subjective definition of hate speech, neglecting user characterization, potential circumvention of policies, and the difficulty of adapting to rapidly evolving hate speech expressions. Similarly, in [25], the authors explored whether a model trained on a dataset from one social media platform unrelated to any specific domain could effectively classify hate speech in the sports domain. The experiments involved Hate Speech detection in Serbian using BiLSTM deep neural networks, resulting in good precision but relatively low Recall. A limitation of this work is its specificity to the sports domain.

Furthermore, the authors in [26] applied ensemble learning, specifically AdaBoostClassifier, for hate speech detection in social media. Results demonstrated improved prediction accuracy. However, limitations include the reliance on accurate preprocessing and the need for substantial training data. As social media evolves with new hate speech forms, ongoing adaptation is required for sustained high accuracy. Additionally, in [27], the authors focused on detecting and visualizing hate speech on cyber-watchdog-based social media. They employed various classification algorithms, including Bidirectional Encoder Representations from Transformers (BERT), Support Vector Machine (SVM), Convolutional Neural Networks (CNN), and Attention-Based Models. The study aimed to develop an approach for detecting and visualizing hate speech attacks on social media, but limitations were identified, such as the need for high accuracy in the classification process due to the diversity of hate speech expressions. The authors in [28] presented a tool that takes an offensive language dataset that only contains Standard American English (SAE) text, and then transforms the SAE text into simulated text with the help of style transfer. The model assesses fairness by comparing predictions on SAE and simulated texts, enabling measurement of underrepresented demographic groups. It facilitates fairness testing on new datasets before investing in human annotation for fine-tuning. Moreover, the limitations include imperfections in the method, emphasizing the need for exploring more sophisticated techniques to ensure that generated sentences contain specific offensive words.

Furthermore, the authors in [29] investigated the use of multi-task learning (MTL) to address the challenges of domain transfer in hate speech detection. It is applying a model trained on one dataset to a different dataset, to address this issue, the authors propose an MTL framework that utilizes sentiment analysis as an auxiliary task alongside the primary task of hate speech detection. Additionally, in [30], the authors suggested a multi-class classification approach that categorizes tweets into three categories: hate speech, offensive language, and neither using the Davidson dataset, they noticed that almost 40% of hate speech is misclassified. Moreover, the authors in [31] utilized convolutional neural networks to capture more complex representations of language, to make them suited for handling the long tail of hate speech. They classified the text into hate and non-hate speech.

In [32], the authors suggested a fuzzy multi-task learning model for hate speech type's detection, which utilized semi-supervised multi-label learning from single-labeled data. The results demonstrated the superiority of the fuzzy approach in terms of embedding features. The fuzzy approach also provided an intensity score for the presence of each type of hate speech in a tweet, facilitating the analysis of correlation between different labels; something that probabilistic approaches lacked. Additionally, in [33], the authors applied extremely weakly-supervised methods for hate speech classification. They conducted a comprehensive dataset and cross-dataset experiments to analyze the transferability of hate speech classification models. Weakly supervised classification exhibited advantages over traditional supervised classification, allowing the application of algorithms across various hate speech datasets and domains, utilization of unlabeled documents without domain mismatch issues, and facilitating cross-dataset comparison through "reannotation" of labeled datasets. However, a limitation of this work was its focus on classifying hate speech categories target groups rather than detecting the presence of hate speech in a post.

Neutrosophic sets [34], introduced as a natural extension of fuzzy logic, offer a more adaptable framework for effectively navigating uncertainty. The following studies contribute substantial insights into the diverse applications and advantages of neutrosophic sets. The work in [35] demonstrated the application of neutrosophic sets in multi-attribute group decision-making, showcasing its ability to handle uncertainty in complex evaluations, specifically in the context of evaluating mathematics teachers. Employing single-valued trapezoidal neutrosophic numbers, the study showcases the versatility and robustness of neutrosophic sets in multi-attribute group decision-making scenarios. Moreover,

the authors in [36] compared the performance of neutrosophic approaches against a spectrum of deep learning models. The findings underscore the efficacy of neutrosophic sets in capturing and managing sentiments, presenting a compelling case for their application in complex analytical tasks. Additionally, the authors in [37] contributed a novel approach to skin cancer classification by leveraging fused deep features within a neutrosophic environment, the study demonstrates the adaptability of neutrosophic sets in enhancing accuracy and reliability in medical diagnostics. Also, the work in [38] introduced neutrosophic sets to enhance decision-making in group scenarios. The study contributes advancements in addressing the intricacies of collective decision-making, further expanding the scope of neutrosophic applications. The work in [39] proposed an image processing method utilizing a generalized linguistic neutrosophic cubic aggregation operator, demonstrating its potential for tackling image processing tasks under uncertainty. These diverse studies showcase the growing interest and potential of neutrosophic sets across various domains. By incorporating uncertainty into decision-making and analysis processes, neutrosophic sets offer a valuable tool for enhancing the accuracy and robustness of complex systems.

Our proposed classification model demonstrates distinct features when compared to existing methods. Notably, it excels in handling uncertainty by quantifying it through the calculation of the difference between predicted truth and false memberships, resulting in indeterminacy membership values. In terms of network architecture, our model innovatively employs two separate Multi-Layer Perceptron (MLP) models dedicated to predicting truth and false membership degrees, diverging from the common practice of using single MLPs in existing methods. The optimization strategy of our model is also unique, integrating the Whale Optimization Algorithm (WOA) and Particle Swarm Optimization (PSO) for weight adjustment. This approach sets it apart from existing methods, which often rely on gradient-based or evolutionary algorithms [16]. Additionally, our model incorporates interval neutrosophic sets [40], providing a more comprehensive representation of uncertainty and enhancing its capability for nuanced classification. This inclusion further underscores the novel aspects of our proposed classification model within the broader landscape of existing methods.

3 Proposed Model

Traditional machine learning models [41] for hate speech detection cannot handle uncertainty and ambiguity in natural language, such as fuzzy logic [42]. Fuzzy logic systems face limitations due to manual design complexities and inherent imprecision, fuzzy logic systems are typically designed by human experts, who must specify the membership functions for the fuzzy sets, leading to inaccuracies in critical applications like hate speech detection. The proposed model incorporates two-step Whale Optimization Algorithm-based Multilayer Perceptron (WOA-MLP) and neutrosophic sets, addressing these challenges. The WOA algorithm [16] explores optimal weights during MLP training, and The PSO algorithm [43] fine-tunes weights, mitigating local minima issues. The hybrid WOA-PSO optimizes the MLP training phase, enhancing predictive accuracy for hate speech multi-classification. The model employs two WOA-PSO-MLPs, predicting truth and false membership degrees, the difference between truth and false membership values is used to estimate indeterminacy membership, forming an interval neutrosophic set. This approach improves optimization robustness, offering better solutions. Fig. 1 illustrates the interconnected elements of the WOA-PSO-MLP technique for hate speech multi-classification. Subsequent sections elaborate on these major model elements.

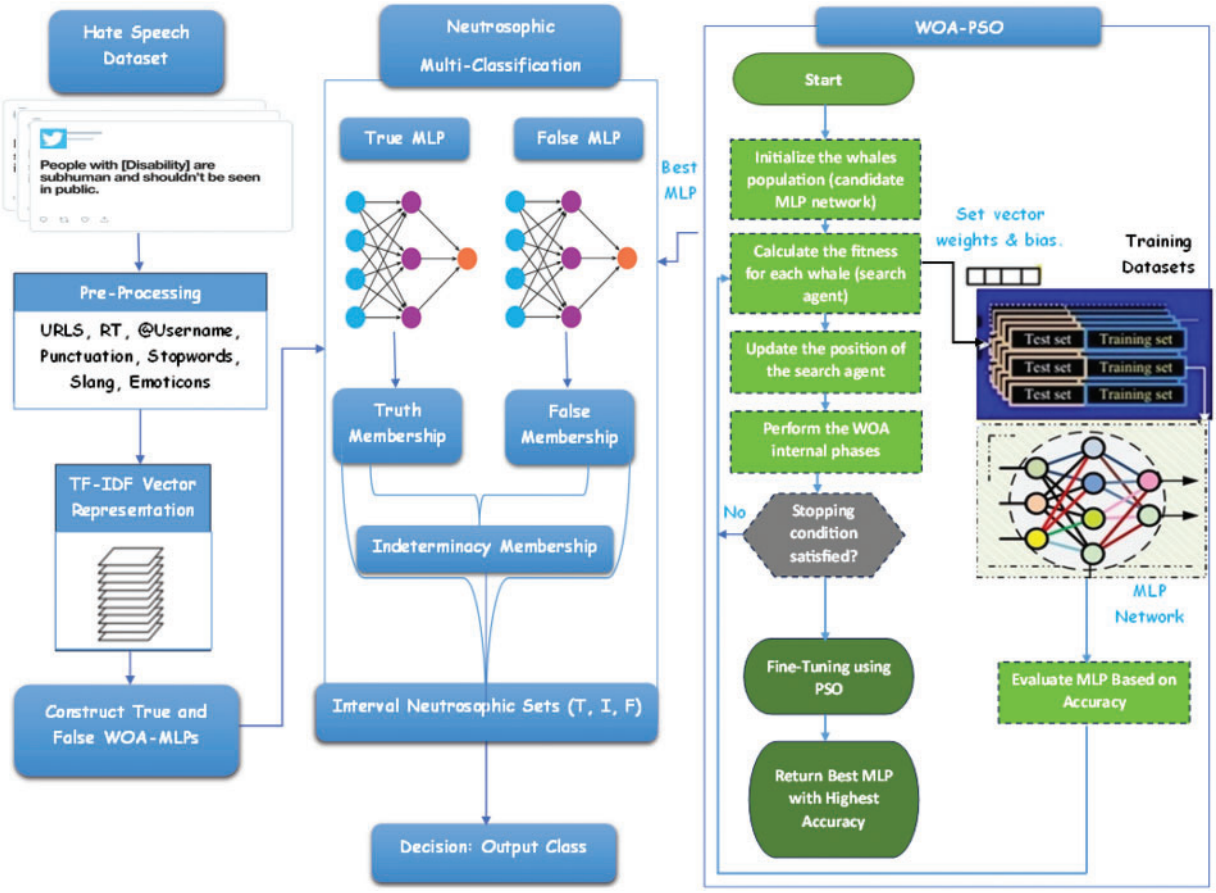

Figure 1: The proposed neutrosophic WOA-PSO-MLP hate speech classification

3.1 Data Collection Phase

This stage focuses on gathering crucial data to validate the proposed WOA-PSO-MLP model, which is applied to Twitter. Twitter serves as an effective platform for data collection on hate speech and offensive speech due to its real-time trend list that reflects current popular topics. However, this application domain presents challenges due to limited context, informal language, abbreviations, and susceptibility to noise, spam, and misleading content. These factors contribute to higher uncertainty in the collected data. Several benchmark datasets exist for hate speech detection, including the Davidson dataset [30]: This dataset contains 24,000 labeled tweets categorized as hate speech, offensive, or neither. It offers a diversity of three types, hate speech, offensive language, and neither. However, it primarily focuses on English language hate speech and may not be representative of other languages or dialects. Twitter Sentiment Analysis dataset [44]: This dataset contains tweets classified as positive, negative, neutral, or irrelevant. While not explicitly focused on hate speech, it provides a broader range of sentiment labels that could be useful for training and validating sentiment-sensitive hate speech detection models. OffensEval dataset [45]: This dataset includes multiple sub-datasets covering various offensive language types, including hate speech, profanity, and cyberbullying. However, the individual sub-datasets are relatively small and may not represent the full spectrum of online offensive content. Hate Speech dataset from white supremacy forum [46]: This dataset contains approximately

16,000 tweets labeled hate speech and not hate speech only, so it is not suitable for multi-classification. The Davidson dataset was chosen for this study due to its representation of hate speech, offensive and neither, its focus on hate speech specifically, and its adequate size for training and validating the proposed model. Also, the dataset primarily focuses on English language hate speech.

3.2 Pre-Processing Phase

Text preprocessing involves removing stop words, URLs, numbers, usernames, and punctuation, as well as replacing slang and emoticons to represent the main body of the text.

3.3 Two-Step WOA for Training MLP Phase

In this section, we describe the proposed approach based on a combination of WOA and Particle Swarm Optimization (PSO) to optimize the training phase typically involves adjusting the weights of the MLP network. The decision to employ WOA in the initial step of our model is grounded in its unique characteristics and proven efficacy in solving optimization problems. WOA is driven by its exploration capabilities as WOA excels at exploring the search space for optimal solutions. Its unique spiral updating mechanism and random search agent selection enable it to discover diverse and potentially optimal solutions, especially when dealing with complex data with multiple local optima. Also, WOA's dynamic update equations allow it to adjust its search behavior throughout the optimization process, overcoming limitations of other algorithms that may stagnate in local optima. This adaptability enhances its ability to navigate complex search landscapes. WOA's relatively simple structure and implementation compared to other advanced metaheuristic algorithms [16].

The goal is to find the optimal set of weights that minimizes the error or loss function during the training process for the MLP by leveraging the exploration capabilities of WOA and the optimization capabilities of PSO. In the proposed approach combining WOA and PSO, the initial solution obtained from the WOA algorithm represents an initial set of weights for the MLP network. The PSO algorithm is then applied to further refine and improve these weights. This refinement process can be seen as a form of fine-tuning, where the PSO algorithm adjusts the weights to optimize the performance of the MLP. In the WOA the search agents are often referred to as whales. The algorithm is inspired by the hunting behavior of humpback whales, where each whale represents a potential solution or candidate solution to the optimization problem. Similar to how whales navigate and search for food in the ocean, the search agents in WOA explore the search space in search of the optimal or near-optimal solution. Each whale (search agent) has a position in the search space, which corresponds to a set of parameters or variables being optimized. The movement and updating of the whales' positions during the optimization process are guided by the WOA algorithm's rules and equations. By simulating the interaction and movement of whales, the WOA algorithm aims to improve the quality of the solutions iteratively. The whales (search agents) communicate and adapt their positions based on their own experiences and the information shared by other whales, leading to a collective improvement in the search for better solutions.

The approach employs a population of search agents, referred to as whales, to identify the global optimum for optimization problems. Similar to other population-based algorithms, the search process initiates by generating a set of random solutions, known as candidate solutions, for a given problem. This set is iteratively improved until a specified end criterion is met. The distinguishing feature of WOA lies in the rules governing the enhancement of candidate solutions at each optimization step. WOA commences optimization by creating a set of random solutions and updates the positions of search agents based on either a randomly selected search agent or the best search agent obtained thus far. It

encourages diverse solutions as WOA promotes diversity within the population of search agents. The algorithm incorporates a spiral updating mechanism that encourages search agents to move around in a circular pattern, which aids in exploring different areas of the search space. This diversity helps prevent premature convergence to suboptimal solutions. WOA employs dynamic update equations that adapt and change over the optimization process. The update equations control the movement and adjustment of search agents' positions, allowing for efficient exploration and exploitation. The dynamic nature of these equations improves the algorithm's ability to adapt to different problems.

WOA-PSO is employed to train a single hidden layer MLP network. When designing this approach, two crucial aspects are considered: the representation of search agents within WOA and the selection of the fitness function. In the Two-Step WOA approach, each search agent is represented as a one-dimensional vector encoding a candidate neural network. The vector consists of weights connecting the input and hidden layers, weights connecting the hidden and output layers, and biases. The vector length corresponds to the total number of weights and biases in the network, and it can be calculated using the following equation:

$$Individual\ length = (n * m) + (2 * m) + 1 \tag{1}$$

where:

n: is the number of input variables that represents the number of features in each text sample.

m: is the number of neurons in the hidden layer, in this work, we follow the same method proposed and used in [47] where m is selected based on the following formula: $2 * N + 1$, where N is the number of dataset features.

To measure the fitness value of the generated two-step WOA agents the fitness value of the generated two-step WOA agents is measured using the accuracy fitness function. The accuracy fitness function is calculated as follows:

$$\text{Accuracy} = (\text{No. of correctly classified test samples})/(\text{Total number of test samples}) \tag{2}$$

The goal of Two-Step WOA using PSO is to find the neural network with the highest accuracy by adjusting the weights of the neural network. The workflow of the Two-Step WOA using the PSO approach applied in this work for training the MLP network can be described in the following steps:

a) Initialization: The initial step involves randomly generating a predetermined number of search agents to represent a potential MLP network. Each search agent is encoded as a one-dimensional vector, with its length corresponding to the total number of weights and biases in the MLP network.

b) Fitness evaluation: Next, the fitness of each search agent (MLP) is evaluated using the accuracy fitness function. The accuracy is determined by calculating the percentage of correctly classified test samples by the neural network. The objective is to find the MLP with the highest accuracy, indicating superior performance.

c) Update the position of search agents: After evaluating the fitness of each search agent, the positions of the search agents are updated using the WOA algorithm. The search agent with the best fitness value, representing the highest performance, is identified, and its weights are stored as the current best solution.

d) Repeat steps b–c: Steps b and c are repeated until the maximum number of iterations is reached, allowing for continuous improvement of the search agents' positions and the corresponding MLP networks.

e) Improve the solution using PSO: The best solution obtained through the WOA algorithm, which represents the MLP network with the highest accuracy, is utilized as the initial set of parameters for the MLP network. The PSO algorithm is then employed to further refine and enhance the solution.

f) Testing: The MLP network with the highest accuracy value is tested on a separate portion of the dataset, consisting of unseen samples, to evaluate its performance in handling new inputs.

Test case and its successful outcomes:

Specific Examples: Hateful Tweet: "You disgusting immigrants are ruining our country! Go back to where you came from!", "Fucking niggers that's all Kentucky is and ever will be" (Correctly classified as hateful), Offensive Tweet: "You're an idiot! I bet you couldn't even tie your shoes!" (Correctly classified as offensive) Neither Tweet: "The weather is beautiful today! I'm going to take a walk in the park". (Correctly classified as neither hate speech nor offensive speech). Handling Edge Cases: Tweet: "Of course, I love everyone equally! Especially those who disagree with me". (Correctly classified as offensive), tweet: "I'm so tolerant, I even tolerate intolerance!" (Correctly classified as neither), Tweet: "I don't like you very much". (Correctly classified as offensive). We included a diverse range of tweets in terms of length, vocabulary, and complexity to assess the model's generalizability.

3.4 Neutrosophic Set Classification Phase

In this paper, we leverage Neutrosophic sets, a mathematical framework introduced by Prof. Florentin Smarandache [34], In the context of a universe Ω, a Neutrosophic set (NS), denoted as A, is defined by three essential membership functions: the truth membership function T_A, the indeterminacy membership function T_A, and the falsity membership function F_A. These functions, represented by real standard elements of the interval [0, 1], characterize the degrees to which an element belongs to the set. Mathematically, a Neutrosophic set A can be expressed in Eq. (3):

$$A = \{< x, (T_A(x), I_A(x), F_A(x)) > x \in E, T_A, I_A, F_A \in]-0, 1+[\} \tag{3}$$

There is no restriction on the sum of $T_A(x)$, $I_A(x)$, and $F_A(x)$ allowing for a flexible representation where the degrees of truth, indeterminacy, and falsity can vary independently. This is mathematically expressed as $0 \leq T_A(x)+I_A(x)+F_A(x) \leq 3$, The unrestricted nature of this sum accommodates a comprehensive characterization of elements within the Neutrosophic set, facilitating a good representation of uncertainty and ambiguity.

An ensemble of two WOA-PSO-MLP is created. The output of the best MLPs is used to represent the membership, non-membership, and indeterminacy of the data points. We use two MLPs to classify data into neutrosophic sets, the output of the MLPs represents the membership, non-membership, and indeterminacy of the data points. Specifically, the first MLP is trained to predict the membership of the data points in the neutrosophic set. The second MLP can be trained to predict the non-membership of the data points in the neutrosophic set. The indeterminacy of the data points can then be calculated using the following Eq. (4):

$$Indeterminacy = 1 - (\text{membership} + \text{non-membership}) \tag{4}$$

The outputs of a single MLP can be modeled using a distributed output code in which each class is assigned a unique codeword, Codewords are unique binary strings that are used to represent classes in a multiclass classification problem. Each class is assigned a unique codeword, and the codewords are designed to be as different from each other as possible to minimize error correlation. One simple way to generate codewords is to use the One-Against-All (OAA) approach [48]. OAA is a multiclass classification technique that uses a single MLP to classify data points into multiple classes. In OAA,

the MLP is trained to predict the probability that a data point belongs to each class. The data point is then assigned to the class with the highest probability. In OAA, each class is assigned a separate codeword, and the codeword for each class is a vector of bits, where the i^{th} bit is equal to 1 if the class is the i^{th} class, and 0 otherwise. To use OAA to generate codewords for a multiclass classification problem, we train a separate MLP for each class. The MLP for the i^{th} class is trained to predict the codeword for the i^{th} class.

The training process involves the Truth MLP, which is designed to estimate the degrees of truth membership. In a k-class truth neural network, the length of a codeword is k, where each bit represents a specific class. For the i^{th} class, the codeword has a value of 1 at the i^{th} position and 0 for the remaining positions. On the other hand, the Falsity MLP shares the same architecture and properties as the Truth MLP but focuses on predicting the degree of false membership. To achieve this, the Falsity MLP is trained using the complement of the target codewords utilized in the training data for the Truth MLP. For instance, if the codeword used to train the Truth MLP for the i^{th} class has a value of 1 at the i^{th} bit and 0 for the rest, then the codeword employed to train the Falsity MLP for the same class will have a value of 0 at the i^{th} bit and 1 for the remaining bits.

When presented with an unknown input pattern, we assign T_j as the true membership of the j^{th} output for the truth network, and F_j as the false membership of the j^{th} output for the falsity network. These predicted outputs are expected to be contradictory, with high truth membership corresponding to low false membership and vice versa. If the truth membership value is high while the false membership value is low, or vice versa, uncertainty arises in predicting these outputs. To quantify the degree of uncertainty or indeterminacy in the prediction, we can calculate the difference between the truth and false membership values. A high difference indicates low uncertainty, whereas a low difference suggests high uncertainty. Let I_j represent the indeterminacy membership of the j^{th} output, which can be calculated as shown in Eq. (5).

$$I_j = 1 - \left|T_j - F_j\right| \tag{5}$$

These three memberships form an interval neutrosophic set. We define A_j as an interval neutrosophic set for the j^{th} output, represented as shown in Eq. (6).

$$A_j = \left\{x(T_{A_j}(x), I_{A_j}(x), F_{A_j}(x))\right\} \tag{6}$$

where T_{A_j} denotes the truth membership function, I_{A_j} represents the indeterminacy membership function, and where F_{A_j} corresponds to the false membership function for the j^{th} output. Interval Neutrosophic Sets (INS) [40] are employed because of their ability to represent the inherent uncertainty in data more effectively than single-valued neutrosophic sets [49]. As data contains inherent ambiguity and vagueness, it is difficult to assign precise membership values using single-valued sets. INS allows representing a range of possible membership degrees, providing a more nuanced and flexible representation of uncertainty. The three memberships form an interval neutrosophic set and are used for decision-making. For each instance, evaluating the INS A_j for all classes. The class with the highest degree of truth $T_{A_j}(x)$ and lowest indeterminacy $I_{A_j}(x)$ is assigned as the final classification.

3.5 Illustrative Example

Here is an illustrative example using tweets for solving neutrosophic classification with three classes: hate speech, offensive language, or neither. We will follow the proposed model's steps:

Truth Membership: Using the first network to predict the degree of truth membership (T_{A_j}) for each class, For the tweet: $T_{A_{hate}} = 0.85$, $T_{A_{offensive}} = 0.1$, and $T_{A_{neither}} = 0.05$.

False Membership: Using the second network to predict the degree of false membership (F_{A_j}) for each class, for the tweet: $F_{A_{hate}} = 0.1$, $F_{A_{offensive}} = 0.8$, and $F_{A_{neither}} = 0.1$.

Uncertainty Quantification: Calculating the indeterminacy membership (I_{A_j}) for each class using the difference between T_{A_j} and F_{A_j}, for the tweet: $I_{A_{hate}} = 1 - \|0.85 - 0.1\| = 1 - 0.75 = 0.25$, $I_{A_{offensive}} = 1 - \|0.1 - 0.8\| = 1 - 0.7 = 0.3$, and $I_{neither} = 1 - \|0.05 - 0.1\| = 1 - 0.05 = 0.95$.

Interval Neutrosophic Set: Forming the interval neutrosophic set A_j for each class using the calculated values: $A_{hate} = \{0.85, 0.25, 0.1\}$, $A_{offensive} = \{0.1, 0.3, 0.8\}$, and $A_{neither} = \{0.05, 0.95, 0.1\}$.

Decision-Making: Evaluate the interval neutrosophic sets for all classes by choosing the class with the highest degree of truth and the lowest indeterminacy as the final classification. In this case, the tweet is classified as "hate speech" due to the highest $T_{A_{hate}}$ and lowest $I_{A_{hate}}$.

4 Discussion

The proposed model is evaluated using the Davidson classification datasets introduced above. In this section, the performance of the proposed model is validated on Davidson's [30] hate speech classification dataset. The Davidson dataset is a dataset of tweets that have been labeled as hateful, offensive, or neither. The dataset was created by collecting tweets that contained words from a hate speech lexicon, and then manually labeling the tweets by CrowdFlower workers. The dataset contains 24,000 tweets, of which 5% are hateful, 15% are offensive, and 80% are neither. The dataset is a valuable resource for training and evaluating machine learning models for hate speech detection and offensive language classification. It is also a useful dataset for researchers studying the nature and prevalence of hate speech and offensive language online. This dataset contains information about tweets and their classifications by CrowdFlower users. Table 1 provides a detailed description of each variable:

Table 1: Davidson dataset description

Variable	Description
Count	No. of CrowdFlower users who coded the tweet (minimum of 3)
Hate speech	No. of CrowdFlower users who classified the tweet as hate speech
Offensive language	No. of CrowdFlower users who classified the tweet as offensive
Neither	No. of CrowdFlower users who considered the tweet neither hate nor offensive
Class	(0: hate speech, 1: offensive language, 2: neither)
Tweet	Text message of the tweet

The experiment was performed in Intel (R), Core (TM) i7 CPU, 8.00 GB RAM implemented in anaconda, using Python. Herein, we utilize the evaluation metrics used in [50]: Precision, Recall, and F1 Score as evaluation metrics [51].

$$Precision = T_P/(T_P + F_P) \tag{7}$$

$$Recall = T_P/(T_P + F_N) \tag{8}$$

$$F1\ Score = 2^*\ (Precision^*Recall)/(Precision + Recall) \tag{9}$$

4.1 Experiment 1: Comparison with Existing Methods

The results validate the effectiveness of the proposed model, demonstrating a 19% increase in accuracy compared to [50] for hate speech detection. One plausible explanation for this outcome lies in the two-step WOA-MLP neutrosophic approach, which enhances classification accuracy. By combining WOA-PSO, the model identifies the optimal set of weights that minimize the error or loss function during MLP training, utilizing the exploration capabilities of WOA and the optimization capabilities of PSO. Then best MLP is used, one MLP for generating the truth membership function and the other MLP is used as the false membership function, and then both are used to calculate indeterminacy membership values as the difference between the truth and false membership values. The results show uncertainty value is equal to 3, so the uncertainty value is low, and it generally indicates that the model's predictions are more reliable and accurate, and the model is more confident in its classifications. Also, this neutrosophic approach handles the ambiguity problems that face hate speech classification, as ambiguity arises when there is a lack of clear distinction between different classes or when it is difficult to confidently assign samples to specific categories. Table 2 shows the classification performance.

Table 2: Comparative study

Ref.	Class	Precision	Recall	F1
Davidson et al. [50]	Hate	0.32	0.53	0.4
	Offensive	0.96	0.88	0.92
	Neither	0.81	0.95	0.87
Talat et al. [29]	Hate	–	–	0.39
	Offensive	–	–	0.94
	Neither	–	–	0.84
Davidson et al. [30]	Hate	0.44	–	–
Zhang et al. [31]	Hate	–	–	0.30
	Non-hate	–	–	0.97
Proposed model	Hate	0.51	0.56	0.54
	Offensive	0.96	0.96	0.96
	Neither	0.86	0.95	0.91

4.2 Experiment 2: Comparison of Model Accuracy with WOA and WOA-PSO

This set of experiments aimed to validate the role of WOA-PSO in enhancing classification accuracy. Table 3 showcases the efficacy of the proposed model, which employs neutrosophic sets and utilizes two WOA-PSO MLPs: one for predicting the degree of truth membership and the other for predicting the degree of false membership. The indeterminacy membership value is calculated by determining the difference between the truth and false membership values for each pair of networks. which raises the accuracy by 3% relative to the model with only WOA-MLP, and 7% relative to the model with only MLP. The experiment demonstrated that incorporating the PSO algorithm in the combined WOA-PSO approach improved the model's performance compared to using WOA alone. By utilizing PSO as a fine-tuning mechanism, the approach enhanced convergence speed and facilitated

a more comprehensive exploration of the search space, ultimately leading to improved classification accuracy.

Table 3: Comparison of precision, recall, and F1 between proposed model, WOA-MLP, and MLP

Ref.	Class	Precision	Recall	F1
Proposed model	Hate	0.51	0.56	0.54
	Offensive	0.96	0.96	0.96
	Neither	0.86	0.95	0.91
WOA-MLP	Hate	0.48	0.53	0.50
	Offensive	0.94	0.94	0.95
	Neither	0.84	0.90	0.88
MLP	Hate	0.44	0.30	0.35
	Offensive	0.94	0.94	0.94
	Neither	0.79	0.88	0.83

4.3 Experiment 3: Comparison between the Proposed Model and Other Nature-Inspired Metaheuristic Algorithms

This set of experiments aimed to compare the efficiency of the combination of WOA-PSO and neutrosophic sets in hate speech classification against a combination of other nature-inspired metaheuristic algorithms with neutrosophic sets. We substituted the WOA module in the proposed model with well-known metaheuristic modules, using their default configurations, through a BlackBox. These metaheuristic algorithms included Ant Colony Optimization (ACO) [52], the Artificial Bee Colony (ABC) Algorithm [53], and the Butterfly Optimization Algorithm (BOA) [54]. The results presented in Table 4 confirm the research hypothesis that integrating WOA-PSO learning optimization improves classification accuracy. The suggested combination achieved a 4% accuracy increase compared to the closest combination involving BOA and the MLP classifier. In ACO, ABC, and BOA, the convergence rate is influenced by key parameters that determine the movement of individuals towards the best position obtained so far by individuals and the group, affecting their tendency to explore or exploit. This, in turn, affects the balance between the exploration and exploitation tendencies of these algorithms. In contrast, the WOA algorithm tends to be more explorative, as evidenced by its lower convergence rate when compared to the other methods. Unlike ACO, ABC, and BOA, the convergence rate in WOA is not significantly influenced by specific parameters within the algorithm. This robustness of WOA makes it a versatile optimization algorithm capable of effectively addressing a wide array of optimization problems. Therefore, the choice of the WOA-PSO combination in this experiment was driven by the aim to leverage the explorative nature of WOA while benefiting from the fine-tuning capabilities of PSO. By incorporating PSO, the optimization process was further enhanced, leading to improved classification accuracy in the task of hate speech classification.

4.4 Experiment 4: Comparison between the Proposed Model, Other Machine Learning Algorithms, and Fuzzy

This experiment aimed to compare the accuracy of the proposed model, machine learning algorithms, and the fuzzy model in hate speech detection. These machine-learning algorithms include

the Support Vector Machine (SVM) [55], the Random Forest (RF) Algorithm [12], and the Logistic Regression (LR) Algorithm [56]. The choice of these machine learning algorithms was based on their popularity and effectiveness in various classification tasks. Each algorithm has its strengths and weaknesses, and the goal was to assess how the proposed neutrosophic model performs in comparison. The results demonstrated that the proposed model achieved a 3% and 6% higher accuracy compared to machine learning algorithms and the fuzzy model, respectively. There are a few reasons that the neutrosophic model achieved higher accuracy: Handling uncertainty and indeterminacy; The neutrosophic model incorporates the concepts of indeterminacy-membership and falsity-membership, allowing it to handle uncertain and conflicting information more effectively. In hate speech classification, where the boundaries between different categories can be blurry, the ability to capture and represent these uncertainties can lead to improved accuracy. Modeling complex relationships; Hate speech classification can involve complex relationships and patterns in the data. The neutrosophic model, combined with the mentioned machine learning algorithms, may have better captured and modeled these complex relationships, leading to improved classification accuracy. accurate understanding of hate speech; The neutrosophic model's consideration of multiple membership degrees allows it to capture the nuances of hate speech more effectively. By incorporating indeterminacy-membership and falsity-membership, the model can recognize and classify instances that may have conflicting or uncertain characteristics, leading to improved accuracy.

Table 4: Results of the comparison between different nature-inspired metaheuristic algorithms

Nature-inspired metaheuristic	Class	Precision	Recall	F1
Proposed model	Hate	0.51	0.56	0.54
	Offensive	0.96	0.96	0.96
	Neither	0.86	0.95	0.91
ACO	Hate	0.47	0.44	0.50
	Offensive	0.90	0.90	0.89
	Neither	0.81	0.84	0.91
ABC	Hate	0.39	0.41	0.35
	Offensive	0.91	0.91	0.87
	Neither	0.80	0.80	0.82
BOA	Hate	0.41	0.40	0.38
	Offensive	0.84	0.86	0.86
	Neither	0.81	0.84	0.82

Also, one key difference between fuzzy sets and neutrosophic sets is their consideration of different membership degrees. Fuzzy sets solely focus on the truth-membership degree, indicating the extent to which an element belongs to a set. On the other hand, neutrosophic sets incorporate additional membership degrees, including indeterminacy-membership and falsity-membership. By considering indeterminacy-membership and falsity-membership, the neutrosophic model can effectively handle situations where there is incomplete or conflicting information. This capability allows the model to detect hate speech more accurately, leading to improved performance. The comparison results in

Table 5 highlight the superiority of the proposed neutrosophic model over the fuzzy model and other machine-learning techniques.

Table 5: Comparison between proposed model, machine learning algorithms, and fuzzy-based model

Algorithm	Class	Precision	Recall	F1
Proposed model	Hate	0.51	0.56	0.54
	Offensive	0.96	0.96	0.96
	Neither	0.86	0.95	0.91
Fuzzy-based model	Hate	0.45	0.27	0.34
	Offensive	0.92	0.97	0.94
	Neither	0.82	0.83	0.82
SVM	Hate	0.43	0.33	0.37
	Offensive	0.93	0.95	0.94
	Neither	0.85	0.87	0.86
Random forest	Hate	0.46	0.06	0.11
	Offensive	0.89	0.98	0.92
	Neither	0.88	0.64	0.74
Logistic regression	Hate	0.48	0.23	0.31
	Offensive	0.93	0.95	0.94
	Neither	0.84	0.88	0.86

5 Conclusions

In this paper, our research addresses the challenging task of hate speech detection in social media. Existing approaches often struggle with the inherent ambiguity and complexity in distinguishing between hateful and offensive content in social media. Our use of neutrosophic logic effectively addresses this challenge by capturing uncertainty and contextually vague content, leading to a more accurate representation of hate speech. This work contributes significantly to the field of social media forensics and offers several notable advantages. The integration of WOA and PSO enables the optimization of MLP models, resulting in enhanced performance and accuracy. The WOA algorithm explores the optimal set of weights during the training process, while the PSO algorithm fine-tunes the weights to further optimize the MLP's performance. This combination of optimization techniques contributes to the overall effectiveness of our model. Additionally, our model predicts degrees of truth membership and false membership, allowing for the estimation of indeterminacy membership or uncertainty in the predictions. This estimation provides valuable insights into the level of uncertainty associated with the classification results, aiding users and decision-makers in assessing the reliability of the outcomes. Furthermore, our research introduces the concept of neutrosophic sets, which represent uncertainty in predictions and resolve ambiguity between true and false classifiers. This technique improves the accuracy and performance of hate speech classification compared to state-of-the-art methods, particularly demonstrated through the evaluation of the Davidson dataset. To further enhance the performance of the proposed model and its applicability in real-world

scenarios, several promising research avenues will be explored in future work. The model will be adapted to support multilingual analysis by incorporating techniques for language-agnostic feature extraction and multilingual embedding models. Dimensionality reduction techniques like principal component analysis and linear discriminant analysis will be employed to reduce the feature space and improve computational efficiency. Ensemble methods incorporating the proposed model with other classification models like support vector machines or random forests will be investigated to leverage the strengths of diverse models and improve generalizability. Parallelization and distributed computing techniques will be explored to leverage the power of multiple processing units and achieve scalable real-time performance. Moreover, other types of neutrosophic sets will be used like Trapezoidal neutrosophic numbers.

Acknowledgement: The authors would like to thank the editors and reviewers for their valuable work, as well as the supervisor and family for their valuable support during the research process.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Yasmine M. Ibrahim, Saad M. Darwish, Reem Essameldin; data collection: Yasmine M. Ibrahim; analysis and interpretation of results: Yasmine M. Ibrahim; draft manuscript preparation: Yasmine M. Ibrahim, Saad M. Darwish, Reem Essameldin. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the author Yasmine M. Ibrahim upon reasonable request.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] A. K. Jain, S. R. Sahoo, and J. Kaubiyal, "Online social networks security and privacy: Comprehensive review and analysis," *Complex Intell. Syst.*, vol. 7, no. 5, pp. 2157–2177, Oct. 2021. doi: 10.1007/s40747-021-00409-7.

[2] N. Alkiviadou, "Hate speech on social media networks: Towards a regulatory framework?" *Inf. Commun. Technol. Law*, vol. 28, no. 1, pp. 19–35, Jul. 2018. doi: 10.1080/13600834.2018.1494417.

[3] Y. Delgado, B. S. Price, P. J. Speaker, and S. L. Stoiloff, "Forensic intelligence: Data analytics as the bridge between forensic science and investigation," *Forensic Sci. Int. Synergy.*, vol. 3, no. 2, pp. 100162, Aug. 2021. doi: 10.1016/j.fsisyn.2021.100162.

[4] A. R. Javed, W. Ahmed, M. Alazab, Z. Jalil, K. Kifayat and T. R. Gadekallu, "A comprehensive survey on computer forensics: State of the art, tools, techniques, challenges, and future directions," *IEEE Access*, vol. 10, pp. 11065–11089, Jan. 2022. doi: 10.1109/ACCESS.2022.3142508.

[5] S. Naqvi *et al.*, "Privacy-preserving social media forensic analysis for preventive policing of online activities," in *10th IFIP Int. Conf. NTMS*, Canary Islands, Spain, Jun. 24–26, 2019, pp. 1–6.

[6] N. M. Zainudin, M. Merabti, and D. L. Jones, "Online social networks as supporting evidence: A digital forensic investigation model and its application design," in *2011 Int. Conf. on Res. and Inno. in Inf. Sys.*, Kuala Lumpur, Malaysia, Nov. 23–24, 2011, pp. 1–6.

[7] R. Montasari and R. Hill, "Next-generation digital forensics: Challenges and future paradigms," in *IEEE 12th ICGS3*, London, UK, Jan. 16–18, 2019, pp. 205–212.

[8] D. Khurana, A. Koli, and K. Khatter, "Natural language processing: State of the art, current trends and challenges," *Multimed. Tools. Appl.*, vol. 82, no. 3, pp. 3713–3744, Jan. 2023. doi: 10.1007/s11042-022-13428-4.

[9] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," *ACM Comput. Surv.*, vol. 51, no. 4, pp. 1–30, Jul. 2018. doi: 10.1145/3232676.

[10] B. Al-Duwairi, A. S. Shatnawi, H. Jaradat, A. Al-Musa, and H. Al-Awadat, "On the digital forensics of social networking web-based applications," in *10th Int. Conf. ISDFS*, Istanbul, Turkey, Jun. 6–7, 2022, pp. 1–6.

[11] G. L. Kovács, P. Alonso, and R. Saini, "Challenges of hate speech detection in social media," *SN Comput. Sci.*, vol. 2, no. 2, pp. 203, Feb. 2021. doi: 10.1007/s42979-021-00457-3.

[12] N. S. Mullah and W. M. N. W. Zainon, "Advances in machine learning algorithms for hate speech detection in social media: A review," *IEEE Access*, vol. 9, pp. 88364–88376, Jun. 2021. doi: 10.1109/ACCESS.2021.3089515.

[13] S. U. Hassan, J. Ahamed, and K. Ahmad, "Analytics of machine learning-based algorithms for text classification," *SUSOC*, vol. 3, pp. 238–248, Apr. 2022. doi: 10.1016/j.susoc.2022.03.001.

[14] S. MacAvaney, H. R. Yao, E. Yang, K. Russell, N. Goharian and O. Frieder, "Hate speech detection: Challenges and solutions," *PLoS One*, vol. 14, no. 8, pp. e0221152, Aug. 2019. doi: 10.1371/journal.pone.0221152.

[15] M. Shyamsunder and K. S. Rao, "Classification of LPI radar signals using multilayer perceptron (MLP) neural networks," in *Proc. ICASPACE*, Singapore, Dec. 2022, pp. 233–248.

[16] H. Nadimi-Shahraki, Z. Zamani, Z. Varzaneh, and S. Mirjalili, "A systematic review of the whale optimization algorithm: Theoretical foundation, improvements, and hybridizations," *Arch. Comput. Methods Eng.*, vol. 30, no. 7, pp. 4113–4159, May 2023. doi: 10.1007/s11831-023-09928-7.

[17] F. M. Plaza-del-arco, D. Nozza, and D. Hovy, "Respectful or toxic? Using zero-shot learning with language models to detect hate speech," in *WOAH*, Toronto, Canada, Jul. 2023, pp. 60–68.

[18] A. Schmidt and W. Michael, "A survey on hate speech detection using natural language processing," in *Proc. 5th Int. Workshop on SocialNLP*, Valencia, Spain, Apr. 2017, pp. 1–10.

[19] V. Christianto and F. Smarandache, "A review of seven applications of neutrosophic logic: In cultural psychology, economics theorizing, conflict resolution, philosophy of science, etc.," *J. Multidiscip. Sci.*, vol. 2, no. 2, pp. 128–137, Mar. 2019. doi: 10.3390/j2020010.

[20] F. Smarandache, "Neutrosophic logic-a generalization of the intuitionistic fuzzy logic," *SSRN*, vol. 4, no. 3, pp. 396, Jan. 2016. doi: 10.2139/ssrn.2721587.

[21] S. Das, B. K. Roy, M. B. Kar, S. Kar, and D. Pamučar, "Neutrosophic fuzzy set and its application in decision making," *J. Ambient Intell. Humaniz. Comput.*, vol. 11, no. 11, pp. 5017–5029, Mar. 2020. doi: 10.1007/s12652-020-01808-3.

[22] T. M. Shami, A. A. El-Saleh, M. Alswaitti, Q. Al-Tashi, M. A. Summakieh and S. Mirjalili, "Particle swarm optimization: A comprehensive survey," *IEEE Access*, vol. 10, pp. 10031–10061, Jan. 2022. doi: 10.1109/ACCESS.2022.3142859.

[23] F. E. Ayo, O. Folorunso, F. T. Ibharalu, I. A. Osinuga, and A. Abayomi-Alli, "A probabilistic clustering model for hate speech classification in Twitter," *Expert Syst. Appl.*, vol. 173, no. 1, pp. 114762, Jul. 2021. doi: 10.1016/j.eswa.2021.114762.

[24] R. Martins, M. Gomes, J. J. Almeida, P. Novais, and P. Henriques, "Hate speech classification in social media using emotional analysis," in *Proc. IEEE BRACIS*, Sao Paulo, Brazil, Oct. 22–25, 2018, pp. 61–66.

[25] S. V. Stanković and M. Mladenović, "An approach to automatic classification of hate speech in sports domain on social media," *J. Big Data*, vol. 10, no. 109, pp. 766–797, Jun. 2023. doi: 10.1186/s40537-023-00766-9.

[26] A. Haque and M. N. U. R. Chowdhury, "Hate speech detection in social media using the ensemble learning technique," *Int. J. Neural Netw. Adv. Appl.*, vol. 15, no. 1, pp. 5815–5821, May 2023. doi: 10.35444/IJANA.2023.15111.

[27] S. Modha, P. Majumder, T. Mandl, and C. Mandalia, "Detecting and visualizing hate speech in social media: A cyber Watchdog for surveillance," *Expert Syst. Appl.*, vol. 161, pp. 113725, Dec. 2020. doi: 10.1016/j.eswa.2020.113725.
[28] A. Rios, "FuzzE: Fuzzy fairness evaluation of offensive language classifiers on African-American english," in *Proc. AAAI Conf.*, Washington DC, USA, Apr. 2020, pp. 881–889.
[29] Z. Talat, J. Thorne, and J. Bingel, "Bridging the gaps: Multi task learning for domain transfer of hate speech detection," in J. Golbeck, Ed., *Online Harassment. Human-Computer Interaction Series*, Cham, Switz: Springer, July 2018, pp. 29–55, Accessed: May 19, 2022. 10.1007/978-3-319-78583-7_3
[30] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proc. 11th ICWSM*, Montreal, Quebec, Canada, May 15–18, 2017, pp. 512–515.
[31] Z. Zhang and L. Luo, "Hate speech detection: A solved problem? the challenging case of long tail on Twitter," *Semant. Web J.*, vol. 10, no. 5, pp. 925–945, Sep. 2019. doi: 10.3233/SW-180338.
[32] P. Liu, H. Burnap, W. Alorainy, and M. L. Williams, "Fuzzy multi-task leaning for hate speech type identification," in *Proc. the WWW '19*, New York, NY, USA, May 2019, pp. 3006–3012.
[33] Y. Jin, L. Wanner, V. L. Kadam, and A. Shvets, "Towards weakly-supervised hate speech classification across datasets," in *Proc. 7th WOAH*, Toronto, Canada, Jul. 2023, pp. 42–59. 10.18653/v1/2023.woah-1.4
[34] F. Smarandache, M. Ali, and M. Khan, "Arithmetic operations of neutrosophic sets, interval neutrosophic sets and rough neutrosophic sets," in *Fuzzy Multi-Criteria Decision-Making Using Neutrosophic Sets*, 1st ed. vol. 369. Cham, Switz: Springer, Nov. 2019, pp. 25–42.
[35] I. Irvanizam and N. Zahara, "An Extended EDAS based on multi-attribute group decision making to evaluate mathematics teachers with single-valued trapezoidal neutrosophic numbers," in S. Broumi, Ed., *Handbook of Research on the Applications of Neutrosophic Sets Theory and their Extensions in Education*, Chocolate Ave. Hershey, PA 17033, USA: IGI Global Press, Jun. 2023, pp. 40–67. 10.4018/978-1-6684-7836-3.ch003
[36] M. Sharma, I. Kandasamy, and W. B. Vasantha, "Comparison of neutrosophic approach to various deep learning models for sentiment analysis," *Knowl.-Based Syst.*, vol. 223, no. 2, pp. 107058, Jul. 2021. doi: 10.1016/j.knosys.2021.107058.
[37] A. Abdelhafeez, H. K. Mohamed, A. Maher, and N. A. Khalil, "A novel approach toward skin cancer classification through fused deep features and neutrosophic environment," *Front. Public Health*, vol. 11, pp. e1265, Apr. 2023. doi: 10.3389/fpubh.2023.1123581.
[38] I. Irvanizam *et al.*, "An improved EDAS method based on bipolar neutrosophic set and its application in group decision-making," *Appl. Comput. Intell. Soft. Comput.*, vol. 2021, no. 3, pp. 1–16, Oct. 2021. doi: 10.1155/2021/1474629.
[39] G. Kaur and H. Garg, "A new method for image processing using generalized linguistic neutrosophic cubic aggregation operator," *Complex Intell. Syst.*, vol. 8, no. 6, pp. 4911–4937, Dec. 2022. doi: 10.1007/s40747-022-00718-5.
[40] H. Wang, D. Madiraju, Y. Q. Zhang, and R. Sunderraman, "Interval neutrosophic sets," *Int. J. Appl. Math. Stat.*, vol. 3, pp. 1–18, Sep. 2004. doi: 10.48550/arXiv.math/0409113.
[41] S. Abro, S. Shaikh, Z. Hussain, Z. Ali, S. U. Khan and G. Mujtaba, "Automatic hate speech detection using machine learning: A comparative study," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 8, pp. 484–491, 2020. doi: 10.14569/issn.2156-5570.
[42] A. Fernández, V. López, M. J. D. Jesus, and F. Herrera, "Revisiting evolutionary fuzzy systems: Taxonomy, applications, new trends and challenges," *Knowl.-Based Syst.*, vol. 80, no. 6, pp. 109–121, May 2015. doi: 10.1016/j.knosys.2015.01.013.
[43] A. G. Gad, "Particle swarm optimization algorithm and its applications: A systematic review," *Arch. Comput. Methods Eng.*, vol. 29, no. 5, pp. 2531–2561, Oct. 2022. doi: 10.1007/s11831-021-09694-4.
[44] S. Hussein, "Twitter sentiments dataset," *Type: Dataset*, pp. 1–3468, May 2021. doi: 10.17632/z9zw7nt5h2.1.

[45] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra and R. Kumar, "SemEval-2019 Task 6: Identifying and categorizing offensive language in social media (OffensEval)," in *Proc. 13th SemEval*, Minneapolis, Minnesota, USA, June 2019, pp. 75–86.
[46] O. de Gibert, N. Pérez, A. García-Pablos, and M. Cuadros, "Hate speech dataset from a white supremacy forum," in *Proc. ALW2*, Brussels, Belgium, Oct. 2018, pp. 11–20.
[47] I. Aljarah, H. Faris, and S. Mirjalili, "Optimizing connection weights in neural networks using the whale optimization algorithm," *Soft. Comput.*, vol. 22, no. 1, pp. 1–15, Jan. 2018. doi: 10.1007/s00500-016-2442-1.
[48] P. Kraipeerapun, C. C. Fung, and K. W. Wong, "Multiclass classification using neural networks and interval neutrosophic sets," in *5th WSEAS*, Venice, Italy , Nov. 20–22, 2006.
[49] H. Zhang, J. Wang, and X. Chen, "Interval neutrosophic sets and their application in multicriteria decision making problems," *Sci. World J.*, vol. 2014, no. 3, pp. 1–15, Feb. 2014. doi: 10.1155/2014/645953.
[50] T. Davidson, D. Bhattacharya, and I. Weber, "Racial bias in hate speech and abusive language detection datasets," in *Proc. 3rd ALW*, Florence, Italy, Aug. 2019, pp. 25–35.
[51] M. Hossin and M. Sulaiman, "A review on evaluation metrics for data classification evaluations," *Int. J. Data Min. Knowl. Manag. Process*, vol. 5, no. 2, pp. 01–11, Mar. 2015. doi: 10.5121/ijdkp.2015.5201.
[52] M. Morin, I. Abi-Zeid, and C. Quimper, "Ant colony optimization for path planning in search and rescue operations," *Eur. J. Oper. Res.*, vol. 305, no. 1, pp. 53–63, Feb. 2023. doi: 10.1016/j.ejor.2022.06.019.
[53] E. Kaya, B. Gorkemli, B. Akay, and D. Karaboga, "A review on the studies employing artificial bee colony algorithm to solve combinatorial optimization problems," *Eng. Appl. Artif. Intell.*, vol. 115, no. 3, pp. 105311, Oct. 2022. doi: 10.1016/j.engappai.2022.105311.
[54] S. Arora and S. Singh, "Butterfly optimization algorithm: A novel approach for global optimization," *Soft Comput.*, vol. 23, no. 3, pp. 715–734, Feb. 2019. doi: 10.1007/s00500-018-3102-4.
[55] N. Sevani, I. A. Soenandi, Adianto, and J. Wijaya, "Detection of hate speech by employing support vector machine with word2Vec model," in *Proc. 7th ICEEIE*, Malang, Indonesia, Oct. 2021, pp. 1–5. doi: 10.1109/ICEEIE52663.2021.9616721.
[56] P. Jemima and P. Preethy, "Hate speech detection using machine learning," in *Proc. 7th ICCES*, Coimbatore, India, Jun. 22–24, 2022, pp. 141–145.