IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 March 2024, accepted 22 April 2024, date of publication 24 April 2024, date of current version 2 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3393295

RESEARCH ARTICLE

Social Media Forensics: An Adaptive Cyberbullying-Related Hate Speech Detection Approach Based on Neural Networks With Uncertainty

YASMINE M. IBRAHIM[1,2], **REEM ESSAMELDIN**[3], **AND SAAD M. SAAD**[1,4]

[1] Department of Information Technology, Institute of Graduate Studies and Research, Alexandria University, Alexandria 21526, Egypt
[2] Faculty of Computers and Information Technology, The Egyptian E-Learning University (EELU), Giza 12611, Egypt
[3] Faculty of Computers and Data Science, Alexandria University, Alexandria 21554, Egypt
[4] Department of Artificial Intelligence, Faculty of Computer Sciences and Artificial Intelligence, Pharos University in Alexandria, Alexandria 21648, Egypt

Corresponding author: Yasmine M. Ibrahim (igsr.yasmine@alexu.edu.eg)

ABSTRACT Cyberbullying is a social media network issue, a global crisis affecting the victims and society. Automatically identifying cyberbullying on social media has become extremely hard because of the complicated nature and intricate language employed within these platforms. The brevity and informal nature of text often results in ambiguous or unclear expressions, making it challenging to accurately interpret the intended meaning. Identifying cyberbullying becomes even more complex when faced with uncertain or contextually vague content. Presently, numerous approaches are available for cyberbullying detection, However, they continue to grapple with the challenge of distinguishing between various forms of cyberbullying-related hate speech due to its ambiguous and vague nature, and they also fall short in terms of accuracy. This paper proposes a novel approach to fine-grained cyberbullying classification by integrating Neutrosophic Logic within the Multi-Layer Perceptron (MLP) model. The proposed model enhances cyberbullying types by mitigating the challenges posed by the ambiguity and overlapping boundaries between distinct categories of cyberbullying. The incorporation of Neutrosophic Logic aims to address the uncertainty, ambiguity, and indeterminacy within classification decisions, offering a more comprehensive and flexible approach for handling complex classification scenarios. The model, leveraging the one-against-one strategy in MLP classification, captures complex relationships between various types of cyberbullying, due to the overlaps and ambiguous instances within cyberbullying types. The testing phase of this model emphasizes the significance of Neutrosophic Logic, employing class probabilities from multiple one-against-one classifiers to provide a comprehensive insight into classification outcomes. The results of the proposed model demonstrate the performance enhancement of incorporating Neutrosophic Logic for fine-grained cyberbullying classification tasks.

INDEX TERMS Cyberbullying, hate speech detection, one-against-one, multiclass classification, neutrosophic sets, social media forensics.

I. INTRODUCTION

With the progression of digital technologies and the widespread adoption of social media, bullying has escalated in its threat to individuals, as it now can be carried out using internet technologies [1]. Threats, online harassment, disgrace, fear, and other forms of cyberbullying are characterized as new forms of violence or bullying that are perpetrated through technical gadgets and the World Wide Web [2]. Social media forensics involves the collection, analysis, and

The associate editor coordinating the review of this manuscript and approving it for publication was Angelo Trotta.

investigation of digital data gathered from diverse social media platforms to uncover evidence pertinent to legal or criminal inquiries. Within the realm of digital forensics, social media evidence represents a novel area of study [3]. Social media evidence analysis plays a crucial role in cyberbullying detection. Cyberbullying detection is a challenging task because the ambiguity of language can vary greatly depending on the speaker, the audience, the context, the informality of language, and the diversity of cultures and contexts [4], [5]. There are two main approaches to cyberbullying speech detection [6]: machine learning-based, and ensemble approach. The machine learning (ML) approach uses statistical models to learn the patterns of language that are associated with cyberbullying related hate speech. Furthermore, ensemble approach combines the machine learning-based approaches. This approach uses ML models to confirm whether the post is considered as cyberbullying speech. This can help to improve the accuracy of cyberbullying speech classification. Within the field of machine learning, the MLP is a widely employed technique [7].

MLP classification is a ML method that can be employed to classify cyberbullying in textual content. MLP classifiers are composed of multiple layers of artificial neurons, which are interconnected in a specific way. The neurons in each layer are responsible for learning different features of the input text, and the output of the final layer is used to classify the input text into different categories, such as cyberbullying or not. The advantages of using MLP is that it can learn complex nonlinear relationships between features, which makes them well-suited for text classification tasks. MLPs can handle large amounts of data, which is often the case with text classification tasks. MLPs are relatively easy to train, which makes them a good choice for tasks where data is limited [7].

The inherent subjectivity of language in verbal communication poses challenges in identifying and categorizing different types of cyberbullying. This complexity arises from the fact that the interpretation of text can vary depending on several factors, such as the context of its usage, the intentions of the speaker or writer, and the cultural background of the audience [8]. A text that may be considered as cyberbullying speech in one context may not be considered cyberbullying in another. Models are typically trained on a dataset of labeled text, but this dataset may not be representative of all the different ways that cyberbullying can be expressed. As a result, machines may sometimes misclassify cyberbullying as non-cyberbullying, or vice versa.

Neutrosophic logic (NL) [9] is an extension of classical logic that introduces a third truth value, besides true and false, to represent indeterminacy. NL allows for the handling of uncertainty and ambiguity in reasoning and decision-making processes. Neutrosophic logic finds applications in various domains including artificial intelligence, decision support systems, and pattern recognition, offering a more comprehensive approach to dealing with imperfect or incomplete information. NL has numerous advantages over traditional classification approaches. First, its ability to represent and reason with indeterminate and vague information. Traditional classification methods often struggle to handle uncertainty in data, leading to inaccurate or incomplete results. NL, however, provides a formal framework for representing and reasoning with uncertain information, allowing for more robust and flexible classification. Another advantage of NL is its ability to capture and model complex relationships between variables in a more nuanced way. Traditional classification approaches may oversimplify or overlook subtle interactions between factors, leading to less accurate classifications.

In contrast, NL employs a three-valued representation to delineate levels of truth, falsity, and indeterminacy. Conversely, deep learning relies on probabilities, while ML typically only accounts for truth and falsity. Fuzzy Logic [10], [11], on the other hand, signifies uncertainty through degrees of membership and non-membership. NL's approach, characterized by its explicit representation of indeterminacy and membership functions, positions it as a more effective tool for addressing the intricacies of detecting cyberbullying-related hate speech compared to traditional fuzzy tools and ML.

A. CONTRIBUTION AND METHODOLOGY

This paper offers a cyberbullying fine grain classification model based on neutrosophic neural networks. The suggested model proposes a novel approach to cyberbullying types of classification using neutrosophic logic. By applying a one-against-one approach for multiclass classification using MLP classifier. Additionally, neutrosophic classification is performed on the probabilities of each class.

The main contributions of this article are: (1) Introducing a new fine grained neutrosophic neural network classification model. (2) Constructing and training an ensemble of binary classifiers to tackle Multi Classification using the One-Against-One strategy. (3) MLP classifier is used to predict class probabilities for cyberbullying types. (4) Generating probabilities for each class using a set of binary classifiers and subsequently extracting the dominant class for the given cyberbullying types using these probabilities. (5) Converting probabilities into neutrosophic sets for final classification decision based on interval neutrosophic sets.

The rest of the paper is organized as follows: Section II presents some of the recent related work. Section III describes the proposed cyberbullying classification model. In Section IV the results, and discussions on the cyberbullying dataset. Finally, conclusions are drawn in Section V.

II. RELATED WORK

Cyberbullying detection has been widely researched, beginning with user studies in the social sciences and psychology sectors, and more recently shifting to computer science with the goal of building models for automated identification. There are many kinds of ML techniques, however, the most well-known and extensively utilized form, supervised ML, was utilized in virtually all research on cyberbullying prediction on social media. Nevertheless, there is no one optimum ML method for all issues. As a result, most study chooses and evaluates a variety of supervised classifiers to find the best fit for their issue. The most widely used predictors in the area, as well as the data attributes accessible for trials, are utilized to pick classifiers. Researchers, on the other hand, may only pick which algorithms to use for building a cyberbullying detection model after conducting a full practical trial [6].

The authors of the work [12] assess ML approaches against the lexical method, recognizing limitation in identifying verbally expressed emotions despite achieving high indicators. To overcome this constraint, the authors propose sentiment identification methods using knowledge bases associated with specific emotions. The study introduces three distinct cyberbullying recognition approaches: a rules-based method that identifies explicit cyberbullying through keyword combinations and lexical resources, supervised machine learning that analyzes various linguistic features, and deep machine learning utilizing neural networks such as convolutional neural networks. Each approach offers unique advantages. The rules-based method provides interpretability for explicit cyberbullying identification, supervised learning allows flexibility with diverse linguistic features, and deep learning captures complex patterns and relationships. However, limitations include the potential oversight of understated cyberbullying types in the rules-based method, the need for substantial labeled data in supervised learning, the computational intensity of deep learning, and potential challenges with very long texts.

Additionally, the authors in [13] suggested an automated cyberbullying detection model to deal with imbalanced short text and diverse dialects appears in the Arabic text. The simulated annealing optimization algorithm is used to find the optimal set of samples from the majority class to balance the training set. The work employed a comprehensive evaluation by testing the model with both traditional machine learning algorithms and deep learning algorithms. This approach ensures a robust assessment of the framework's performance across different methodologies. The authors mentioned that the limitation of this work is associated with the complexities introduced by linguistic diversity and regional variations in the Arabic language, particularly when applied to cyberbullying detection.

Furthermore, the authors in [14] introduced a strategy for social media cyberbullying detection. They employed four machine learning models: Support Vector Machine (SVM), Naïve Baise (NB), Decision Tree (DT), and K-Nearest Neighbor (KNN) to categorize texts into cyberbullying and non-cyberbullying categories. The training of these models involved the application of various features, including bad words, negative emotion, positive emotion, links, proper nouns, and pronouns. The work didn't deal with cyberbullying sub-types. Similarly, the authors in [15] developed an ensemble model for cyberbullying detection. They used Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN), which have demonstrated effectiveness in detecting cyberbullying. The results demonstrate the method's efficacy in identifying and categorizing offensive language on social media platforms. The authors admit there's still a lot of work to do in making more dependable methods for spotting cyberbullying. They point out challenges, especially the difficulty of making the method work well in different situations. They suggest looking into advanced techniques and new ways of using technology to improve cyberbullying detection, emphasizing the need for ongoing innovation.

In [16] the authors deployed three deep and six learning algorithms for cyberbullying classification. The results show that LSTM is the highest method for cyberbullying detection in terms of accuracy and recall. But they didn't deal with class imbalance data, fine grain classification to classify types of cyberbullying. Additionally, the work in [17] developed a framework to determine cyberbullying in texts, the framework employs a Fuzzy Logic System that uses the outputs of SVM classifiers as its inputs to identify the cyberbullying. Results show that it is necessary to improve the accuracy of SVM classifiers to determine the bullying severity through Fuzzy Logic. Also, the limitation of the work is the challenge in figuring out how severe instances of bullying were based on the collected tweets. Despite using a fuzzy logic system, the authors found it tough to consistently identify the severity of bullying episodes. They discovered that determining how severe a bullying episode was become difficult because each author had a different view, even when they used the same criteria to create the fuzzy rules. This means that the authors didn't always agree on how serious a bullying situation was, making it a subjective and challenging aspect of their research.

Neutrosophic sets [18], introduced as an extension of fuzzy logic, present a more versatile approach for successfully handling uncertainty. These studies provide significant understandings into the different applications and benefits of neutrosophic sets. The authors outlined in [19] the utilization of neutrosophic sets in multi-attribute group decision-making, highlighting their capability to handle uncertainty in intricate assessments, particularly when evaluating mathematics teachers. By engaging single-valued trapezoidal neutrosophic numbers, the study underscores the adaptability and resilience of neutrosophic sets in scenarios involving multi-attribute group decision-making. Furthermore, the authors in [20] established an innovative approach for skin cancer classification, utilizing fused deep features within a neutrosophic framework. This study shows how neutrosophic sets enhance accuracy and reliability in medical diagnostics, showcasing their adaptability in this environment. Moreover, the work presented in [21] introduced an image processing procedure utilizing a generalized linguistic neutrosophic cubic aggregation operator, highlighting its effectiveness in addressing image processing challenges during uncertainty.

These varied studies underscore the increasing interest and promise of neutrosophic sets in diverse domains. By integrating uncertainty into decision-making and analytical procedures, neutrosophic sets emerge as a valuable tool for improving the precision and resilience of complex systems.

The conducted survey showed that the cyberbullying social media detection systems have the following limitations: (a) Handling Class Imbalance: Many cyberbullying datasets suffer from class imbalance, where the number of instances of cyberbullying types may be significantly lower than other cyberbullying types of instances. Failure to address this issue can lead to biased models and decreased performance in detecting cyberbullying accurately.

(b) Fine-Grained Classification: While some works focus on binary classification of cyberbullying versus non-cyberbullying, there's a need for more fine-grained classification to differentiate between various types or severity levels of cyberbullying. Ignoring this aspect may lead to oversimplified models that cannot effectively address the degrees of cyberbullying behavior.

(c) Subjective Determination of Bullying: The subjective process of assessing the severity of bullying incidents presents a hurdle in reliably identifying and classifying instances of cyberbullying. This subjectivity can result in discrepancies in data labeling and model evaluation, ultimately affecting the dependability of cyberbullying detection systems.

(d) Fuzzy approach is dealing with uncertainty, but it has some disadvantages that can make it unsuitable in fine grained cyberbulying as fuzzy logic systems are typically designed by human experts, who must specify the membership functions for the fuzzy sets. This can be a time-consuming and error-prone process. Also, fuzzy logic systems are based on fuzzy sets, which are inherently imprecise. This can lead to inaccurate results, especially in applications where high accuracy is like cyberbullying detection.

(e) Inaccurate classification results are found when using MLP classification method separately, whereas combining it with neutrosophic improves the classification results. To the best of our knowledge, little attention has been paid to devising a new neutrosophic technique for cyberbullying fine-grained classification.

III. METHODOLOGY

In order to make accurate fine-grained classifications for cyberbullying types, initially, the model utilizes an MLP classifier employing the One-Against-One strategy, enabling it to discern intricate patterns in the data. Subsequently, probabilities for each cyberbullying class are extracted through this process, providing a rich understanding of the likelihood of different types of cyberbullying occurrences. These probabilities are then converted into neutrosophic sets, leveraging the flexibility and adaptability of neutrosophic logic to capture uncertainties and complexities in the classification task. Finally, utilizing neutrosophic intervals, the model makes the ultimate classification decision, offering a refined approach to cyberbullying detection that accounts for the inherent ambiguities and intricacies of online communication. Fig. 1 illustrates the key components of the model and their interconnected relationships. Subsequent sections elaborate on these major model elements.

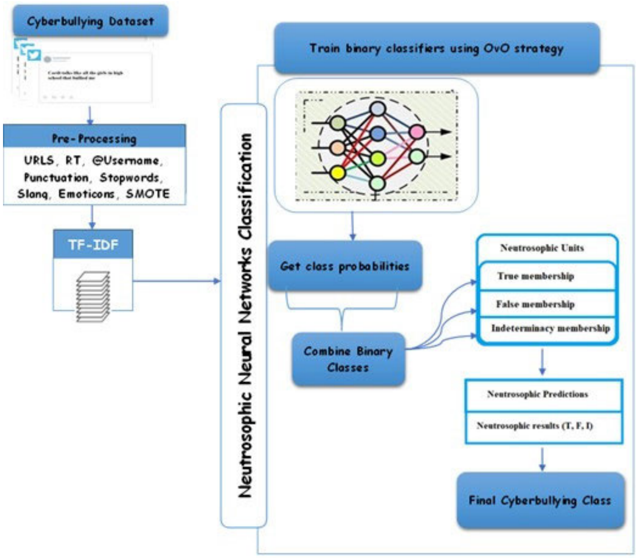

Cyberbullying Dataset
Pre-Processing
URLS, RT, @Username, Punctuation, Stopwords, Slang, Emoticons, SMOTE
TF-IDF
Neutrosophic Neural Networks Classification
Train binary classifiers using OvO strategy
Get class probabilities
Combine Binary Classes
Neutrosophic Units
True membership
False membership
Indeterminacy membership
Neutrosophic Predictions
Neutrosophic results (T, F, I)
Final Cyberbullying Class

FIGURE 1. The proposed Neutrosophic cyberbullying fine-grained classification.

A. DATA COLLECTION PHASE

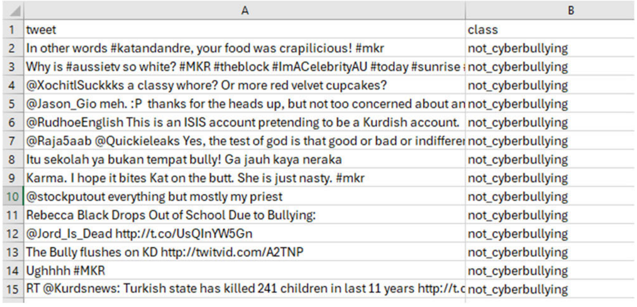
This phase is concerned with collecting the data essential for validating the proposed neutrosophic neural network model. Twitter was chosen to apply the model. There is a cyberbullying dataset [22], this dataset contains more than 47000 tweets labelled according to the class of cyberbullying that contains cyberbullying, gender, other_cyberbullying, age, not_cyberbullying, and ethnicity. The dataset contains two columns: (tweet_text, cyberbullying_type), 'tweet_text' contains tweets. cyberbullying_type contains four types of cyberbullying, age, gender, ethnicity, religion in addition to other_cyberbullying column and not cyberbullying. Fig. 2 shows a sample of the cyberbullying dataset.

B. PRE-PROCESSING PHASE

In text preprocessing, texts are cleared by stripping emoji from text, removing stop words, remove punctuations, links, mentions and new line characters, clear hashtag and special characters to represent the main body of the text.

C. FINE-GRAINED CLASSIFICATION

Fine-grained classification is a more challenging ML type where the goal is to predict the specific subcategory or class within a larger category. Fine-grained cyberbullying classification is the task of classifying cyberbullying incidents into specific subcategories like cyberbullying, age, gender, ethnicity, etc. Fine-grained classification, one-against-one (OvO)



	A	B
1	tweet	class
2	In other words #katandandre, your food was crapilicious! #mkr	not_cyberbullying
3	Why is #aussietv so white? #MKR #theblock #ImACelebrityAU #today #sunrise	not_cyberbullying
4	@XochitlSuckkks a classy whore? Or more red velvet cupcakes?	not_cyberbullying
5	@Jason_Gio meh. :P thanks for the heads up, but not too concerned about an	not_cyberbullying
6	@RudhoeEnglish This is an ISIS account pretending to be a Kurdish account.	not_cyberbullying
7	@Raja5aab @Quickieleaks Yes, the test of god is that good or bad or indifferer	not_cyberbullying
8	Itu sekolah ya bukan tempat bully! Ga jauh kaya neraka	not_cyberbullying
9	Karma. I hope it bites Kat on the butt. She is just nasty. #mkr	not_cyberbullying
10	@stockputout everything but mostly my priest	not_cyberbullying
11	Rebecca Black Drops Out of School Due to Bullying:	not_cyberbullying
12	@Jord_Is_Dead http://t.co/UsQInYW5Gn	not_cyberbullying
13	The Bully flushes on KD http://twitvid.com/A2TNP	not_cyberbullying
14	Ughhhh #MKR	not_cyberbullying
15	RT @Kurdsnews: Turkish state has killed 241 children in last 11 years http://t.c	not_cyberbullying

FIGURE 2. Cyberbullying dataset sample.

[23], and one-against-all (OvA) [24] classification are related techniques that can be used to improve the accuracy of ML models for text classification tasks. In multiclass classification, the goal is to classify text into one of multiple categories. One-against-one classification works by training a separate binary classifier for each pair of classes. For example, if there are four classes, then six binary classifiers would be trained. Each binary classifier would be trained to distinguish between one pair of classes. To classify a new text instance, all six binary classifiers would be used to predict the probability that the instance belongs to each class. The class with the highest predicted probability is assigned to the instance.

One-against-all classification works by training a separate binary classifier for each class against the rest of the classes. For example, if there are four classes, then four binary classifiers would be trained. Each binary classifier would be trained to distinguish between one class and the rest of the classes. To classify a new text instance, all four binary classifiers would be used to predict the probability that the instance belongs to each class. The class with the highest predicted probability is assigned to the instance [22].

D. BINARY CLASSIFICATION

The binary classification problem aims to find a linear function able to correctly classify an input vector between two classes. Given a training set $Z = (x_i, y_i){:}i \in 1, \dots, l$ with points $x_i \in R^d$ and classes $y_i \in -1, +1\}$, where Z^+ is $(x_i, y_i) \in Z{:}y_i = +1\}$, the positive class set, and Z^- is $(x_i, y_i) \in Z{:}y_i = -1\}$, the negative class set, the objective is to utilize an MLP to delineate complex decision boundaries between these classes represented by a normal vector $w \in R^d$ and a bias $b \in R$. The training process involves adjusting the weights and biases iteratively based on the error observed in the training set. The objective is to find the optimal weights (w) and bias (b) that minimize the classification error [25].

E. MULTI-CLASS CLASSIFICATION

In many real-world applications, a classifier must be able to classify an input vector between n classes, n $\in$ N. Given a training set $Z = (x_i, y_i){:}i \in 1, \dots, l$ with points $x_i \in R^d$ and classes , $y_i \in 1, 2, \dots, n$, the objective is to build a function capable of assigning the correct class to an input vector. The optimization problem that emerges from the expansion of the original binary formulation for large margin classifiers to work with more than two classes becomes highly complex according to the increase of the number of classes. Solving a binary classification problem is faster than solving a multiclass classification with the same amount of data. Therefore, instead of expanding the formulation of the binary classification, it is more common to break the multiclass classification problem into binary ones and combine the answers from the binary classifiers to assign the correct class [23]. In this section, we present the two main approaches for solving multiclass classification: one-against-one and one-against-all [24].

1) ONE-AGAINST-ALL

The one-against-all approach takes into consideration each class j against the others, where $j \in 1, 2, \dots, n$, for breaking the multiclass problem into a binary classification problem. For each class j, the full training set Z is taken into consideration, but the class j is seen as the positive class and the other classes are seen as the negative class. A decision boundary $(w, b)_j$ is then generated for each class j following a MLP and stored in a decision boundary set H. At the end of the process, n decision boundaries are generated, where n is the number of classes. Each decision boundary tells whether an input is likely to be of class j or not. The final class of an input is decided by finding out which decision boundary is the closest to the input. The class related to this decision boundary is then assigned to the input. Decision boundary equation (w_j, b_j) for each class j: $w_j^T x + b_j = 0$, The number of decision boundaries grows linearly with regard to the number of classes [23], [24].

2) ONE-AGAINST-ONE

The one-against-one approach takes into consideration pairs of classes (j, k), where $j, k \in 1, 2, \dots, n$ and $j < k$, for breaking the multiclass into a binary classification. For each pair (j, k), a subset Z of the original training set Z consisting of points with classes j and k is created, where j can be seen as the positive class and k can be seen as the negative class. The subset Z is used to generate a decision boundary $(w, b)_{j,k}$ following MLP, and the decision boundary is added to the decision boundary set H. This process generates in total $n(n-1)/2$ decision boundaries, where n is the number of classes. A decision boundary $(w, b)_{j,k}$ predicts the class j if the input is classified as positive and k otherwise. A new input must be classified by every decision boundary and the class with the highest predicted probability is finally assigned to the input. The decision boundary (w, b) for each pair (j, k) can be represented by the equation: $\mathrm{w}^T_{(j,k)}x + b_{(j,k)} = 0$, where w^T is the transpose of the weight vector, x is the input instance, and b is the bias term. The sign of this equation determines the classification result for class j and class k. The number of decision boundaries grows quadratically with regard to the number of classes [23], [24].

3) NEUTROSOPHIC CLASSIFICATION

The concept of NL is applied in the fine-grained classification process described in Fig. 1. NL [9] allows for the representation of uncertainty, ambiguity, and indeterminacy in classification decisions, providing a more comprehensive and flexible approach to handle complex classification scenarios where crisp boundaries between classes may not exist. In the proposed model, the neutrosophic aspect is introduced through the utilization of the OvO approach with multiple binary MLP classifiers. In OvO approach, each binary classifier is trained to distinguish between a specific pair of classes, capturing the relationship and nuances between them. This approach acknowledges and accounts for the possibility of overlapping or ambiguous instances that may fall between two classes, a common occurrence in fine-grained classification tasks. During the training phase, the binary MLP classifiers are trained on the dataset using the OvO strategy [24]. This strategy involves creating a separate classifier for each pair of classes by selecting samples and labels corresponding to those classes. By training multiple binary classifiers, the model gains a deeper understanding of the intricate boundaries and relationships between the classes. This approach enables the model to capture the complex decision boundaries necessary for fine-grained classification tasks. In the testing phase, the neutrosophic concept is further emphasized. For each instance in the testing set, predictions are obtained from all the one-against-one classifiers.

During the testing phase, the predicted probabilities [26] for each class can be obtained by evaluating the input instance x with each decision boundary $(w, b)_{j,k}$ and applying a softmax function, given by Eq. 1:

$$
\begin{aligned}
P(j|x) = \exp(-w\_((j,k))\wedge Tx + [\![b]\!]\_((j,k)))/ \\
\times ((\exp(-w\_((j,k))\wedge Tx + [\![b]\!]\_((j,k))) \\
+ \exp(w\_((j,k))\wedge Tx + [\![b]\!]\_((j,k))))) \quad (1)
\end{aligned}
$$

where $P(j|x)$ and represent the probabilities of the input instance x belonging to class j and class k, respectively.

The predicted probabilities from each classifier are collected and sorted in descending order. By considering the collective knowledge from all the classifiers, the model can make more informed decisions, taking into account the uncertainty and ambiguity associated with each instance's class assignment. The class with the highest predicted probability is assigned to the instance, representing the most likely class membership, considering the input's neutrosophic nature. By incorporating the concept of neutrosophic logic, the fine-grained classification model becomes more robust and capable of handling complex classification scenarios. It allows for the representation of uncertainty and ambiguity, enabling the model to make nuanced decisions even in situations where crisp class boundaries do not exist. This approach enhances the model's accuracy and performance by considering the relationships between classes and capturing the inherent uncertainty present in fine-grained classification tasks.

Converting neutrosophic [27]: converting class probabilities to Neutrosophic Sets (NS): $N(P) = (T, I, F)$, Truth-Membership (T) represents the degree to which the sample belongs to the class. We set a threshold T and assign T if $P \geq$T, and 0 otherwise. Indeterminacy-Membership (I) represents the degree of uncertainty or ambiguity. We set a threshold I and assign I if $I \leq P < T$, and 0 otherwise. Falsity-Membership (F) represents the degree to which the sample does not belong to the class. You can assign F if $P < I$, and 0 otherwise.

Final Classification Decision based on Interval Neutrosophic Sets (INS) [28]: INS are an extension of traditional NS that provide a more flexible representation of uncertainty. In INS instead of specifying precise values for the degrees of truth, indeterminacy, and falsity, Intervals is defined for these parameters. These intervals allow for a range of possible values, capturing the inherent uncertainty and imprecision in the data more effectively. The use of INS offers several advantages including decision-making, classification, and risk assessment.

IV. ILLUSTRATIVE EXAMPLE

In this section, an illustrative example is provided for solving the proposed neutrosophic classification with five classes: Age, Ethnicity, Gender, Religion and Other types of cyberbullying.

Tweet text example = "Hey loser, why don't you go cry to your mommy? You're pathetic". Following the proposed model's steps:

Step 1: Tokenizing the tweet into individual words or tokens.

Tokens: ["Hey", "loser", "why", "don't", "you", "go", "cry", "to", "your", "mommy", "You're", "pathetic"].

Step 2: Converting the tokens into numerical vectors using TF-IDF technique.

TF-IDF assigns weights to each token based on its frequency in the tweet and rarity across the dataset. Each token is represented by a vector of numerical values.

TF-IDF values for each token): "Hey": [0.1, 0.0, 0.05, 0.0, ..., 0.02], "loser": [0.0, 0.2, 0.0, 0.0, ..., 0.03], "why": [0.05, 0.0, 0.08, 0.0, ..., 0.0], "don't": [0.0, 0.0, 0.0, 0.1, ..., 0.0], ...

Step 3: One-vs-One Classification using MLP.

After obtaining the TF-IDF vectors, we feed them into multiple MLP classifiers trained for each pair of classes. We train multiple MLP classifiers, each focusing on distinguishing between a pair of classes. These classifiers are trained using the one-vs-one strategy. Each MLP classifier takes the TF-IDF vectors as input and produces predictions for each class pair. For example, we have classes Age, Ethnicity, Gender, Religion, and Other_cyberbullying, we train classifiers for pairs like (Age vs. Ethnicity), (Age vs. Gender), (Age vs. Religion), (Age vs. Other_cyberbullying), (Ethnicity vs. Gender), and so on.

Classifier for (Age vs. Ethnicity) predicts Age: 0.2, Ethnicity: 0.1, Classifier for (Age vs. Gender) predicts Age: 0.1, Gender: 0.3, Classifier for (Age vs. Religion) predicts Age: 0.4, Religion: 0.5, Classifier for (Age vs. Other) predicts Age: 0.2, Other: 0.4, Classifier for (Ethnicity vs. Gender) predicts Ethnicity: 0.2, Gender: 0.3.

Step 4: Extracting probabilities for each class pair.

For example: Probability of Age: 0.9, Probability of Ethnicity: 0.3, Probability of Gender: 0.3, Probability of Religion: 0.5, Probability of Other_Cyberbullying: 0.4.

Step 5: Fine-grained Classification:

Converting probabilities to Neutrosophic Sets: converting the probabilities into neutrosophic sets for each class using specified thresholds (T, I, F). Assuming $T = 0.9, I = 0.3, F = 0.2$:

Example (for Age):
Given probability for Age: 0.9 (since 0.9 >= T)
T(Age) = 0.9
I(Age) = |0.9 - 0.5| = 0.4 (since T > P(Age) >= F)
F(Age) = 1 - 0.9 = 0.1 (since P(Age) < F)

Example (for Ethnicity):
Given probability for Ethnicity: 0.3 (F < 0.3 < T)
T(Ethnicity) = 0.0
I(Ethnicity) = |0.3 - 0.5| = 0.2 (since T > P(Ethnicity) >= F)
F(Ethnicity) = 1 - 0.3 = 0.7 (since P(Ethnicity) < T)

Example (for Gender):
Given probability for Gender: 0.3 (F < 0.3 < T)
T(Gender) = 0.0
I(Gender) = |0.3 - 0.5| = 0.2 (since T > P(Gender) >= F)
F(Gender) = 1 - 0.3 = 0.7 (since P(Gender) < T)

Example (for Religion):
Given probability for Religion: 0.5 (F < 0.5 < T)
T(Religion) = 0.0
I(Religion) = |0.5 - 0.5| = 0.0 (since P(Religion) = T)
F(Religion) = 1 - 0.5 = 0.5 (since P(Religion) < T)

Example (for Other):
Given probability for Other_Cyberbullying: 0.4 (F < 0.4 < T)
T(Other_Cyberbullying) = 0.0
I (Other_ Cyberbullying) = |0.4 - 0.5| = 0.1,
(since T > P(Other) >= F)
F (Other_ Cyberbullying) = 1 - 0.4 = 0.6 (since P(Other) < T)

Step 6: Final Classification Decision Using Interval Neutrosophic Set.

Interval Neutrosophic Set for: Age: 0.9, 0, 0.1}, Ethnicity: 0.0, 0.2, 0.7}, Gender: 0.0, 0.2, 0.7}, Religion: 0.0, 0, 0.5}, and Other_Cyberbullying: 0.0, 0.1, 0.6}. These interval neutrosophic sets represent the truth, indeterminacy, and falsity memberships for each class, calculated based on the given probabilities and thresholds. for Age, the truth membership (T) is highest (0.9) among all classes, and the indeterminacy membership (I) is also relatively low (0.4). Therefore, according to the neutrosophic classification, the final decision is to classify the tweet as related to "Age." In this case, the tweet is classified as Cyberbullying_type: Age.

V. EXPERIMENTAL RESULTS

In this section, the performance of the proposed model is validated on two datasets focused on cyberbullying classification in social media, an arena where its prevalence and impact have grown considerably. The experiment was conducted using an Intel (R) Core (TM) i3 processor with 8.00 GB RAM and implemented in Anaconda. Herein, we utilize the evaluation metrics used in [6]: Precision, Recall, and F1 Score as evaluation metrics [19].

$$Precision = T_P / (T_P + F_P) \tag{2}$$

$$Recall = T_P / (T_P + F_N) \tag{3}$$

$$F1Score = 2*(Precision*Recall)/(Precision + Recall) \tag{4}$$

Cyberbullying dataset 1 [22] contains over 47,000 labeled tweets specifically classified according to various categories related to cyberbullying: Age, Ethnicity, Gender, Religion, Other types of cyberbullying, and not classified as cyberbullying.

Cyberbullying dataset 2 [29] contains over total of approximately 100,000 tweets classified according to many categories related to cyberbullying: Race/Ethnicity, Gender/ Sexual, Religion, Other types of cyberbullying, and not_cyberbullying.

A. EXPERIMENT 1: MODEL PERFORMANCE FOR FINE GRAINED CYBERBULLYING CLASSIFICATION

In this experiment we dropped the (not_cyberbullying) data form classification type in datasets to make a fine grain classification between cyberbullying types: age, gender, ethnicity, religion, and other cyberbullying. We got 95% accuracy of our proposed model using dataset 1 and 97 % using dataset 2. We defined thresholds (T, I, F) for each class probability and convert to neutrosophic sets; $T = 0.9$ (Truth threshold), $I = 0.3$ (Indeterminacy threshold), $F = 0.2$ (Falsity threshold). Table 1 shows the neutrosophic classification report, The reason of this result is because of the combination of extensive text cleaning operations, including removing emojis, handling contractions, eliminating punctuation and non-ASCII characters, along with handling URLs and mentions, significantly refines the data. These steps are crucial to ensure uniformity, relevance, and consistency within the dataset, thus enhancing the model's ability to extract meaningful patterns from text data.

Also, using SMOTE [30] plays a pivotal role in addressing the class imbalance problem by artificially generating synthetic instances for the minority class. This technique essentially bridges the gap between classes by oversampling the underrepresented class, thus avoiding bias towards the majority class. By creating synthetic examples of the minority class, SMOTE prevents the model from favoring the more dominant class and allows it to learn effectively from both classes. Consequently, this leads to a more balanced and representative learning process, culminating in a model that better generalizes over both major and minor classes.

TABLE 1. Evaluation results of fine-grained cyberbullying classification.

Data	Class	Precision	Recall	F1
Dataset 1 [22]	age	0.98	0.99	0.98
	ethnicity	0.99	0.98	0.98
	gender	0.89	0.92	0.91
	Other_cyberbullying	0.88	0.87	0.87
	religion	0.99	0.97	0.98
	Accuracy			0.95
Dataset 2 [29]	ethnicity/race	0.99	0.99	0.99
	gender/sexual	0.99	0.98	0.98
	religion	0.89	0.92	0.91
	Accuracy			0.97

TABLE 2. Comparison results of different machine learning methods on cyberbullying dataset.

Algorithm	Class	Precision	Recall	F1
RF [32]	age	0.96	0.98	0.97
	ethnicity	0.98	0.97	0.98
	gender	0.92	0.84	0.88
	Other_cyberbullying	0.78	0.90	0.84
	religion	0.98	0.93	0.96
	Accuracy			0.92
LR [33]	age	0.99	0.96	0.97
	ethnicity	0.95	0.97	0.96
	gender	0.97	0.74	0.84
	Other_cyberbullying	0.71	0.94	0.81
	religion	096	0.90	0.93
	Accuracy			0.90
SVM [31]	age	0.98	0.98	0.98
	ethnicity	0.98	0.98	0.98
	gender	0.83	0.81	0.82
	Other_cyberbullying	0.77	0.82	0.79
	religion	0.99	0.96	0.97
	Accuracy			0.91

The utilization of the one-vs-one strategy with MLP classifier in our model played a pivotal role in achieving the high accuracy observed in our results. By training multiple MLP classifiers, each focusing on distinguishing between a pair of classes, we were able to capture intricate relationships and nuances between different cyberbullying types. This approach allowed the model to learn discriminative patterns specific to each class pair, leading to more precise and refined classification decisions. Furthermore, the extraction of probabilities from the predictions of the classifiers provided valuable insights into the model's confidence levels for each class. These probabilities served as the basis for converting the classification outputs into neutrosophic sets, which enabled a more representation of uncertainty and ambiguity in the classification process.The conversion of probabilities to neutrosophic sets using predefined thresholds (T, I, F) further enhanced the model's ability to handle uncertainty and imprecision inherent in cyberbullying classification tasks. By setting appropriate thresholds for truth, indeterminacy, and falsity memberships, we ensured that the model could make informed decisions while considering the inherent uncertainty in the data.

Moreover, the combination of the one-vs-one strategy with MLP classifier and the conversion to neutrosophic sets allowed our model to effectively navigate the complexities of cyberbullying classification. The one-vs-one strategy provided a robust framework for capturing fine-grained distinctions between different cyberbullying types, while the conversion to neutrosophic sets facilitated a more flexible and good representation of classification outputs. Finally, the comprehensive approach employed in our model, which integrates advanced machine learning techniques with neutrosophic logic principles, contributed to the observed high accuracy in cyberbullying classification. By leveraging the strengths of both methodologies, our model demonstrated a superior ability to handle uncertainty, ambiguity, and overlapping features inherent in cyberbullying data, resulting in precise and reliable classification results.

B. EXPERIMENT 2: COMPARISON BETWEEN THE PROPOSED MODEL AND OTHER MACHINE LEARNING MODELS

This group of experiments was carried out to compare the efficiency of the proposed model and machine learning algorithms in the field of fine-grained cyberbullying classification with a combination of some machine learning algorithms. These machine learning algorithms include Support Vector Machine (SVM) [31], the Random Forest (RF) Algorithm [32], and the Logistic Regression (LR) Algorithm [33]. 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 presented in Table 2 confirmed that the proposed neutrosophic model outperformed the other machine learning algorithms in terms of classification accuracy. The suggested combination achieved a 3% increase in accurately classifying cyberbullying types.

There are a few reasons that neutrosophic model achieved higher accuracy compared to the other algorithms: 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. Modeling complex relationships; cyberbullying fine-grained 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. The neutrosophic model's consideration of multiple membership degrees allows it to distinguish between the types of cyberbullying 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.

Furthermore, SVM, known for its robustness in linear classification tasks, may falter when confronted with the nonlinear intricacies inherent in cyberbullying text data. Conversely, our model, harnessing the power of OvO strategy and MLP classifiers, excels in capturing nonlinear patterns and subtle linguistic cues, leading to more accurate

TABLE 3. Comparison accuracy with different activation functions.

Activation Function	Accuracy
ReLU	95
Tanh	92
Sigmoid	92

classification outcomes. Similarly, while RF exhibits prowess in capturing complex relationships within data, it may succumb to bias towards majority classes in imbalanced datasets. In contrast, our model's utilization of interval neutrosophic sets ensures balanced representation and robust decision-making across all cyberbullying types, thereby mitigating RF's limitations. Furthermore, LR's simplicity and efficiency notwithstanding, its linear nature may limit its ability to capture intricate feature relationships. In contrast, our model's integration of MLP classifiers enables it to learn intricate patterns, effectively overcoming LR's limitations and achieving higher accuracy. Thus, the proposed model's amalgamation of OvO strategy, MLP classifiers, and interval neutrosophic sets addresses the shortcomings of traditional machine learning algorithms, demonstrating superior performance in fine-grained cyberbullying classification tasks.

C. EXPERIMENT 3: THE EFFECT OF USING DIFFERENT ACTIVATION FUNCTION FOR MLP USING CYBERBULLYING DATASET

This set of experiments was performed to compare the accuracy of the proposed model that employs different activation functions like ReLU (Rectified Linear Unit), Sigmoid (Logistic), and Tanh (Hyperbolic Tangent) [34]. The results shown in Table 3 revealed that the use of ReLU activation function improve of 3% for the same method with other activation functions. The performance improvement comes from many factors; non-saturation of gradients: Unlike activation functions like sigmoid or tanh, ReLU does not saturate in the positive region, allowing the gradient to flow smoothly during backpropagation. This facilitates faster convergence during training. Also, Sigmoid and tanh functions suffer from vanishing gradient problems for extremely large or small input values, which can hinder learning in deeper networks. ReLU helps mitigate this issue.

D. EXPERIMENT 4: COMPARISON BETWEEN THE PROPOSED MODEL AND FUZZY SETS USING CYBERBULLYING DATASET

This set of experiments was performed to compare the efficiency of the proposed model and fuzzy logic in the field of fine-grained cyberbullying classification. We apply threshold $= 0.7$ for fuzzy classification. The results presented in Table 4 confirmed that the proposed neutrosophic model outperformed the fuzzy logic in terms of classification accuracy. The suggested model achieved a 2% increase in accurately classifying cyberbullying types. The justification of this result is that fuzzy sets use a single membership grade to handle uncertainty, while neutrosophic sets use three independent membership grades (truth, indeterminacy, and falsity) to provide a more comprehensive representation of uncertainty, especially in situations where truth and falsity are not mutually exclusive and there is room for indeterminacy or ambiguity.

TABLE 4. Evaluation results of fine-grained cyberbullying classification after using fuzzy sets.

Class	Precision	Recall	F1
age	0.98	0.97	0.98
ethnicity	0.99	0.97	0.98
gender	0.90	0.87	0.89
Other_cyberbullying	0.82	0.89	0.85
religion	0.99	0.96	0.98
Accuracy			0.93

TABLE 5. Evaluation results of fine-grained cyberbullying classification after using Bert.

Class	Precision	Recall	F1
age	0.99	0.98	0.98
ethnicity	0.98	0.99	0.98
gender	0.90	0.92	0.91
Other_cyberbullying	0.89	0.88	0.88
religion	0.99	0.97	0.98
Accuracy			0.96

E. EXPERIMENT 5: BERT INTEGRATION IN PREPROCESSING FOR THE PROPOSED MODEL USING CYBERBULLYING DATASET

This experiment was conducted to explore the effectiveness of integrating BERT [15] (Bidirectional Encoder Representations from Transformers) into the preprocessing pipeline for cyberbullying detection. BERT, known for its exceptional language understanding capabilities, was incorporated to enhance contextual analysis of speech and capture nuanced changes in keyword meanings, thereby improving the overall detection accuracy. We applied BERT as part of the text preprocessing step before feeding the data into the classification model. BERT was utilized to tokenize and encode the input text data, ensuring that the semantic context and word meanings were preserved effectively.

Table 5 confirm that the results of this experiment have provided valuable insights into the efficacy of BERT in enhancing the performance of cyberbullying detection models and shed light on the potential benefits of leveraging advanced NLP techniques in this domain. The results of this experiment were indeed promising. The integration of BERT led to a noticeable increase in classification accuracy compared to previous experiments. This improvement can be attributed to several factors, including BERT's ability to capture semantic representations of text, its contextual understanding of language nuances, and its robustness to noise and variations in language usage. Overall, the inclusion of BERT in the preprocessing pipeline represents a significant enhancement to the proposed model's performance, aligning with the modern trends in NLP-based approaches for cyberbullying detection.

TABLE 6. Evaluation results of fine-grained cyberbullying classification after using data augmentation.

Class	Precision	Recall	F1
age	0.99	0.98	0.99
ethnicity	0.98	0.99	0.99
gender	0.79	0.96	0.96
Other_cyberbullying	0.95	096	0.96
religion	0.99	0.99	0.99
Accuracy			0.98

F. EXPERIMENT 6: ENHANCING CYBERBULLYING DETECTION THROUGH DATA AUGMENTATION

This experiment aimed to assess the impact of data augmentation techniques on the performance of cyberbullying detection models. Data augmentation [35] is a prevalent approach used to address dataset insufficiency, especially in scenarios where the available dataset size may be insufficient for large-scale evaluation. We employed various data augmentation techniques to augment the original cyberbullying dataset. These techniques included synonym replacement, random insertion, and random deletion of words within the text samples. The augmented dataset was then combined with the original dataset to create a larger, augmented dataset for training and evaluation. Table 6 shows that the results of the experiment demonstrated the data augmentation techniques significantly improved the performance of cyberbullying detection model. The augmented dataset led to an increase in classification accuracy from 95% to 98%, indicating the effectiveness of data augmentation in mitigating dataset insufficiency issues.

The improvement in results after applying data augmentation techniques can be attributed to three key factors. Firstly, the augmented dataset size provided the model with a more extensive and diverse set of examples to learn from, enhancing its ability to generalize and capture complex patterns. Secondly, by introducing variations in the training data, the model was exposed to a wider range of linguistic scenarios, leading to improved generalization to unseen instances. Lastly, data augmentation helped address class imbalance issues by generating additional samples, ensuring a more balanced representation of minority classes, and thereby improving overall predictive performance.

VI. CONCLUSION AND FUTURE WORK

This paper suggested an accurate model for fine-grained cyberbullying classification. The proposed model uses the integration of NL within the MLP classification model and offers an innovative approach toward handling fine-grained classification scenarios. NL allows the representation of uncertainty, ambiguity, and indeterminacy within classification decisions, thereby enhancing the model's ability to handle complex classification tasks where clear boundaries between classes might be lacking. In this work, we successfully incorporated the principles of Neutrosophic Logic through the utilization of the one-against-one strategy in the training phase. The model, built upon a series of binary MLP classifiers, each discriminating between specific pairs of classes, effectively captured the intricate relationships and nuances between different classes. This approach acknowledges and accounts for potential overlapping or ambiguous instances, addressing the common challenge of intricate class boundaries in fine-grained classification tasks. During the testing phase, the significance of the Neutrosophic concept became further pronounced. The predictions from multiple one-against-one classifiers collectively provided a comprehensive insight into classification outcomes. The extracted dominant class from the Neutrosophic class probabilities showcased the adaptability of the model in handling complex classification scenarios. The results, as evidenced in the comparative analysis of the accuracy between the traditional MLP and the Neutrosophic-empowered MLP, demonstrated the utility and potential performance enhancements offered by incorporating Neutrosophic Logic in the classification process. The model, leveraging Neutrosophic Logic, stands as a flexible and comprehensive solution for fine-grained classification tasks, fostering a deeper understanding of intricate boundaries and relationships between different classes. Future work includes using different languages like Arabic and utilizing GPU with deep learning techniques to discover and enhance the model accuracy. We have also planned to explore the integration of Large Language Models (LLMs) in our future work.

REFERENCES

[1] J. R. W. Yarbrough, K. Sell, A. Weiss, and L. R. Salazar, "Cyberbullying and the faculty victim experience: Perceptions and outcomes," *Int. J. Bullying Prevention*, vol. 5, no. 2, pp. 1–5, Jun. 2023, doi: 10.1007/s42380-023-00173-x.

[2] A. Bussu, S.-A. Ashton, M. Pulina, and M. Mangiarulo, "An explorative qualitative study of cyberbullying and cyberstalking in a higher education community," *Crime Prevention Community Saf.*, vol. 25, no. 4, pp. 359–385, Oct. 2023, doi: 10.1057/s41300-023-00186-0.

[3] 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.

[4] G. Fulantelli, D. Taibi, L. Scifo, V. Schwarze, and S. C. Eimler, "Cyberbullying and cyberhate as two interlinked instances of cyber-aggression in adolescence: A systematic review," *Frontiers Psychol.*, vol. 13, May 2022, Art. no. 909299, doi: 10.3389/fpsyg.2022.909299.

[5] M. S. Jahan and M. Oussalah, "A systematic review of hate speech automatic detection using natural language processing," *Neurocomputing*, vol. 546, Aug. 2023, Art. no. 126232, doi: 10.1016/j.neucom.2023.126232.

[6] G. Kovács, P. Alonso, and R. Saini, "Challenges of hate speech detection in social media," *Social Netw. Comput. Sci.*, vol. 2, no. 2, pp. 1–15, Feb. 2021, doi: 10.1007/s42979-021-00457-3.

[7] 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.

[8] 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 *Proc. 7th WOAH*, Toronto, ON, Canada, Jul. 2023, pp. 60–68.

[9] 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. Res.*, vol. 2, no. 2, pp. 128–137, Mar. 2019, doi: 10.3390/j2020010.

[10] F. Smarandache, "Neutrosophic logic—A generalization of the intuitionistic fuzzy logic," *SSRN Electron. J.*, vol. 4, p. 396, Jan. 2016, doi: 10.2139/ssrn.2721587.

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

[12] R. Zhao and K. Mao, "Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder," *IEEE Trans. Affect. Comput.*, vol. 8, no. 3, pp. 328–339, Jul. 2017.

[13] M. Alzaqebah, G. M. Jaradat, D. Nassan, R. Alnasser, M. K. Alsmadi, I. Almarashdeh, S. Jawarneh, M. Alwohaibi, N. A. Al-Mulla, N. Alshehab, and S. Alkhushayni, "Cyberbullying detection framework for short and imbalanced Arabic datasets," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 35, no. 8, Sep. 2023, Art. no. 101652, doi: 10.1016/j.jksuci.2023.101652.

[14] L. J. Thun, P. L. Teh, and C.-B. Cheng, "CyberAid: Are your children safe from cyberbullying?" *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 7, pp. 4099–4108, Jul. 2022, doi: 10.1016/j.jksuci.2021.03.001.

[15] A. Muneer, A. Alwadain, M. G. Ragab, and A. Alqushaibi, "Cyberbullying detection on social media using stacking ensemble learning and enhanced BERT," *Information*, vol. 14, no. 8, p. 467, Aug. 2023, doi: 10.3390/info14080467.

[16] D. Sultan, A. Toktarova, A. Zhumadillayeva, S. Aldeshov, S. Mussiraliyeva, G. Beissenova, A. Tursynbayev, G. Baenova, and A. Imanbayeva, "Cyberbullying-related hate speech detection using shallow-to-deep learning," *Comput., Mater. Continua*, vol. 74, no. 1, pp. 2115–2131, Apr. 2023, doi: 10.32604/cmc.2023.032993.

[17] C. R. Sedano, E. L. Ursini, and P. S. Martins, "A bullying-severity identifier framework based on machine learning and fuzzy logic," in *Artificial Intelligence and Soft Computing*, vol. 10245, 1st ed. Cham, Switzerland: Springer, 2017, pp. 315–324, doi: 10.1007/978-3-319-59063-9_28.

[18] 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*, vol. 3, 1st ed. Cham, Switzerland: Springer, 2019, ch. 2, pp. 25–42, doi: 10.1007/978-3-030-00045-5_2.

[19] 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 *Handbook of Research on the Applications of Neutrosophic Sets Theory and Their Extensions in Education*, S. Broumi, Ed. Hershey, PA, USA: IGI Global, Jun. 2023, pp. 40–67, doi: 10.4018/978-1-6684-7836-3.ch003.

[20] 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," *Frontiers Public Health*, vol. 11, pp. 1–15, Apr. 2023, doi: 10.3389/fpubh.2023.1123581.

[21] 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.

[22] J. Wang, K. Fu, and C.-T. Lu, "SOSNet: A graph convolutional network approach to fine-grained cyberbullying detection," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Atlanta, GA, USA, Dec. 2020, pp. 1699–1708.

[23] S. Kang, S. Cho, and P. Kang, "Constructing a multi-class classifier using one-against-one approach with different binary classifiers," *Neurocomputing*, vol. 149, pp. 677–682, Feb. 2015, doi: 10.1016/j.neucom.2014.08.006.

[24] W. A. Silva and S. M. Villela, "Improving the one-against-all binary approach for multiclass classification using balancing techniques," *Int. J. Speech Technol.*, vol. 51, no. 1, pp. 396–415, Aug. 2020, doi: 10.1007/s10489-020-01805-1.

[25] W. Wang, L. Feng, Y. Jiang, G. Niu, M.-L. Zhang, and M. Sugiyama, "Binary classification with confidence difference," 2023, *arXiv:2310.05632*.

[26] J. Ma, T. Li, X. Li, S. Zhou, C. Ma, D. Wei, and K. Dai, "A probability prediction method for the classification of surrounding rock quality of tunnels with incomplete data using Bayesian networks," *Sci. Rep.*, vol. 12, no. 1, p. 19846, Nov. 2022, doi: 10.1038/s41598-022-19301-6.

[27] R. Essameldin, A. A. Ismail, and S. M. Darwish, "An opinion mining approach to handle perspectivism and ambiguity: Moving toward neutrosophic logic," *IEEE Access*, vol. 10, pp. 63314–63328, 2022, doi: 10.1109/ACCESS.2022.3183108.

[28] H. Wang, P. Madiraju, Y. Zhang, and R. Sunderraman, "Interval neutrosophic sets," 2004, .

[29] M. Ahmadinejad, N. Shahriar, L. Fan. (2023). *A Balanced Multi-Labeled Dataset for Cyberbully Detection in Social Media*. [Online]. Available: https://www.kaggle.com/datasets/momo12341234/cyberbully-detection-dataset/data

[30] D. Elreedy, A. F. Atiya, and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," *Mach. Learn.*, vol. 2023, pp. 1–21, Jan. 2023, doi: 10.1007/s10994-022-06296-4.

[31] N. M. G. Dwi Purnamasari, M. A. Fauzi, I. Indriati, and L. S. Dewi, "Cyberbullying identification in Twitter using support vector machine and information gain based feature selection," *Indonesian J. Electr. Eng. Comput. Sci.*, vol. 18, no. 3, p. 1494, Jun. 2020, doi: 10.11591/ijeecs.v18.i3.pp1494-1500.

[32] N. Novalita, A. Herdiani, I. Lukmana, and D. Puspandari, "Cyberbullying identification on Twitter using random forest classifier," *J. Phys., Conf. Ser.*, vol. 1192, Mar. 2019, Art. no. 012029, Art. no. 012029, doi: 10.1088/1742-6596/1192/1/012029.

[33] J. M. Ortiz-Marcos, M. Tomé-Fernández, and C. Fernández-Leyva, "Cyberbullying analysis in intercultural educational environments using binary logistic regressions," *Future Internet*, vol. 13, no. 1, p. 15, Jan. 2021, doi: 10.3390/fi13010015.

[34] A. Apicella, F. Donnarumma, F. Isgrò, and R. Prevete, "A survey on modern trainable activation functions," *Neural Netw.*, vol. 138, pp. 14–32, Jun. 2021, doi: 10.1016/j.neunet.2021.01.026.

[35] A. Mumuni and F. Mumuni, "Data augmentation: A comprehensive survey of modern approaches," *Array*, vol. 16, Dec. 2022, Art. no. 100258, doi: 10.1016/j.array.2022.100258.

YASMINE M. IBRAHIM received the B.Sc. degree in computer sciences from the Faculty of Computers and Artificial Intelligence, Helwan University, Egypt, in 2010, and the M.Sc. degree in information technology from the Department of Information Technology, Institute of Graduate Studies and Research (IGSR), Alexandria University, in 2021. She is currently a Researcher and an Assistant Teacher with the Faculty of Computers and Information Technology, The Egyptian E-Learning University. Her research interests include medical image processing, machine learning, and social media forensics.

REEM ESSAMELDIN received the B.Sc. degree in electrical engineering (communications and electronics section) from the Faculty of Engineering, Alexandria University, Egypt, in 2012, the Master of Business Administration (M.B.A.) degree from Zhejiang Normal University, China, in 2016, for a thesis in electronic human resource management (e-HRM), and the M.Sc. degree in information technology from the Department of Information Technology, Institute of Graduate Studies and Research (IGSR), Alexandria University, in 2019. She is currently a Teacher with IGSR. Her research interests include machine learning, social network analysis, opinion mining, e-marketing, business intelligence, e-commerce, and communication networks.

SAAD M. SAAD received the B.Sc. degree in statistics and computer science from the Faculty of Science, Alexandria University, Egypt, in 1995, the M.Sc. degree in information technology from the Department of Information Technology, Institute of Graduate Studies and Research (IGSR), Alexandria University, in 2002, and the Ph.D. degree from Alexandria University, for a thesis in image mining and image description technologies. Since June 2017, he has been a Professor with the Department of Information Technology, IGSR. He is the author or coauthor of more than 50 papers publications in prestigious journals and top international conferences and also received several citations. He has supervised around 80 M.Sc. and Ph.D. students. His research interests include image processing, optimization techniques, security technologies, database management, machine learning, biometrics, digital forensics, and bioinformatics. He has served as a reviewer for several international journals and conferences.