International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Blocking Abuse Comments in Social Network Using Selection Set Algorithm

K. Jothi[1], K. G. Arunkumar [2]

Submitted: 03/10/2024 **Revised:** 20/11/2024 **Accepted:** 02/12/2024

Abstract: Blocking abusive comments on social media is a pressing issue that impacts both individuals and society at large. The challenge of automatically identifying abusive content has become increasingly difficult due to the nuanced language and informal communication styles prevalent on these platforms. The brevity and casual nature of posts often lead to ambiguous expressions, complicating the interpretation of intent. This issue is further exacerbated by the presence of uncertain or contextually vague content. While various methods exist for detecting abusive comments, they often struggle to differentiate between different types of hate speech due to their ambiguous characteristics, resulting in lower accuracy. This paper presents a novel approach for blocking abusive comments by employing a Selection Set Algorithm integrated with a Multi-Layer Perceptron (MLP) model. This approach enhances the classification of abusive comment types by addressing the challenges posed by ambiguity and the overlapping boundaries of different categories. The Selection Set Algorithm is designed to manage uncertainty and vagueness in classification decisions, offering a more robust framework for dealing with complex scenarios. The MLP model, utilizing a one-against-one classification strategy, captures intricate relationships among various types of abusive comments, effectively addressing the overlaps and ambiguities present. The evaluation of this model highlights the effectiveness of the Selection Set Algorithm, employing class probabilities from multiple classifiers to yield comprehensive insights into classification results. The findings indicate a significant performance improvement in blocking abusive comments through the proposed approach.

Keywords*: abuse comments, hate speech detection, one-against-one, multiclass classification, set algorithm*

INTRODUCTION

With the evolution of digital innovations and the pervasive use of social media, the menace of bullying has intensified, as it can now manifest through online platforms [1]. hreats, cyber harassment, humiliation, anxiety, and various forms of digital bullying are recognized as contemporary manifestations of aggression or bullying that are enacted via electronic devices and the internet [2]. Social media forensics entails the gathering, scrutiny, and examination of digital information sourced from a variety of social media outlets to reveal evidence relevant to legal or criminal investigations. Within the scope of digital forensics, the study of social media evidence emerges as a cutting-edge field [3]. Analyzing social media evidence is vital for identifying instances of abuse comments. Detecting abuse comments proves to be a complex endeavor, as the subtlety of language can differ significantly based on the speaker, the receiver, the situation, the informality of expression, and the cultural diversity involved [4], [5]. There are two primary methodologies for detecting abuse comments language [6]: the machine learning-based approach and the ensemble method. The machine learning (ML) technique harnesses the power of statistical frameworks to identify linguistic trends linked to abuse comments-related hate speech. Additionally, the ensemble method integrates various machine learning techniques, using ML models to verify whether a post qualifies as abuse comments discourse.

MLP classification is a machine learning strategy employed for categorizing abuse comments within textual material. MLP classifiers consist composed of numerous tiers of interlinked synthetic neurons, structured in a particular layout. Neurons in each tier are assigned the responsibility of discerning unique characteristics of the input text, with the output from the concluding tier employed to classify the text into different categories, such as harmful remarks or others. The advantages of employing MLP encompass its

PG Scholar1, Assistant Professor2
Department of Computer Science and Engineering, Excel Engineering College, Namakkal, Tamil Nadu 637303
Mailid:1jothirmn@gmail.com,
2kgarunkumar.eec@excelcolleges.com

ability to comprehend intricate nonlinear relationships among features, making it particularly effective for text classification endeavors. MLPs are adept at managing extensive datasets, a frequent occurrence in text classification scenarios. Moreover, MLPs are relatively straightforward to train, rendering them an excellent option for cases where data availability is limited [7].

The subjectivity rooted in language during verbal exchanges creates hurdles in pinpointing and classifying the various forms of abuse comments. This intricacy stems from the reality that text interpretation can differ based on numerous elements, such as the situational context, the aim of the communicator, and the cultural perspective of the audience[8]. An expression deemed as abuse comments in one scenario might not be viewed the same way in another. Typically, models are developed using a labeled text dataset, but this dataset may not encompass all the myriad expressions of abuse comments. Consequently, machines can occasionally misidentify instances of abuse comments as benign or the other way around.

Neutrosophic logic (NL) [9] expands upon classical logic by introducing a third truth value, in addition to true and false, to signify ambiguity. NL accommodates the handling of uncertainty and vagueness in reasoning and decision-making. Neutrosophic logic is utilized across various fields, including artificial intelligence, decision support systems, and pattern recognition, presenting a more holistic method for managing imperfect or incomplete data. NL boasts numerous benefits over conventional classification techniques. Primarily, it can represent and reason with ambiguous and indeterminate information. Traditional classification strategies often falter when faced with uncertainty in data, resulting in flawed or partial outcomes. NL, on the contrary, supplies a structured framework for handling uncertain information, enabling more resilient and adaptable classification. Another benefit of NL is its capability to capture and depict intricate relationships among variables with greater nuance. Conventional classification methods may simplify or neglect subtle interconnections between factors, yielding less precise classifications.

In contrast, NL utilizes a three-valued framework to differentiate levels of truth, falsity, and ambiguity. On the other hand, deep learning is grounded in probabilities, while machine learning typically recognizes only truth and falsehood. Fuzzy Logic [10], [11] signifies ambiguity navigated via fluctuating levels of belonging and exclusion. The approach of NL, distinguished by its vivid portrayal of uncertainty and belonging dynamics, positions it as asuperior instrument for tackling the complexities of identifying abuse comments-related hate speech compared to traditional fuzzy approaches and machine learning.

CONTRIBUTION AND METHODOLOGY

This manuscript unveils a nuanced classification model for abuse comments, utilizing selection set algorithm. The innovative model introduces a fresh perspective on categorizing abuse comments types through the lens of neutrosophic logic. It employs a one-against-one methodology for multiclass classification, leveraging the capabilities of an MLP classifier. Furthermore, the classification is executed on the likelihoods associated with each category. The key contributions of this study include: (1) The introduction of a cutting-edge, finely-tuned selection set algorithm. (2) The development and training of a collection of binary classifiers to address Multiclass Classification via the One-Against-One tactic. (3) The application of the MLP classifier to forecast class probabilities for various abuse comments categories. (4) The generation of probabilities for each category through a suite of binary classifiers, leading to the identification of the predominant class for the respective abuse comments types based on these probabilities. (5) The transformation of probabilities into neutrosophic sets to finalize the classification verdict grounded in interval neutrosophic sets. The layout the structure of the document unfolds as such: Section II delves into contemporary pertinent studies. Section III expands upon theproposed abuse comments classification model. In Section IV, results and discussions surrounding the abuse comments dataset are presented. Ultimately, the conclusions are encapsulated in Section V.

RELATED WORK

The exploration of abuse comments detection has seen extensive investigation, starting from user studies in the realms of human behavior studies and psychological theories, and more recently evolving into the realm of computer science with the purpose of developing frameworks for automated identification. A plethora of machine learning techniques exists, yet the most recognized and frequently applied type, supervised machine learning, has been employed in nearly all studies concerning abuse comments prediction on social platforms. Nonetheless, no single machine learning technique serves as the ultimate solution for every problem. Consequently, most research selects and assesses an array of guided classifiers is employed to identify the optimal one for their unique

dilemma. The frequently applied predictors in this domain, along with the data attributes available for experimentation, serve as the basis for classifier selection. Researchers, however, must conduct comprehensive practical experiments before deciding which algorithms to implement for developing a abuse comments detection model [6].

In their research [12], the authors evaluate machine learning strategies against the lexical approach, acknowledging the limitations in recognizing emotionally charged expressions, despite achieving commendable metrics. To address this limitation, the authors advocate for sentiment analysis techniques leveraging knowledge bases linked to specific feelings. The study puts forth three unique abuse comments detection methodologies: a rules-based technique that identifies overt abuse comments through combinations of keywords and lexical tools, supervised machine learning that scrutinizes various linguistic features, and profound algorithmic training utilizing brain-inspired architectures such as convolutional neural networks. Each methodology presents distinct advantages. The rules-based technique offers clarity for identifying explicit abuse comments, supervised learning provides adaptability with varied linguistic features, and deep learning discerns intricate patterns and connections. However, challenges arise, including the risk of missing subtle forms of abuse comments in the rules-based approach, the necessity for extensive labeled data in supervised learning, the heavy computational demands of deep learning, and potential difficulties with very lengthy texts.

Moreover, the researchers in [13] proposed an automated model for detecting abuse comments to address the challenges posed by uneven short text representation and the variety of dialects present in Arabic. They utilized a simulated annealing optimization algorithm to select the optimal samples from the more prevalent class, ensuring the training set is balanced. The study conducted a thorough evaluation by applying both traditional innovative algorithms in machine intelligence and profound learning approaches to the framework. This tacticguarantees a solid evaluation of the framework's efficacy across various techniques. The authors noted that a significant limitation of their research stems from the complexities arising from linguistic diversity and regional differences within the Arabic language, especially in the context of abuse comments detection.

In addition, the researchers in [14] unveiled a technique aimed at identifying abuse comments on social media platforms. They utilized four machine learning models: The Support Vector Machine (SVM) is a remarkably potent supervised learning technique employed for tasks involving classification and regression. Naïve Bayes (NB) operates as a probabilistic classifier, leveraging Bayes' theorem while making strong independence assumptions. The Decision Tree (DT) employs a tree-like diagram or decision-making framework to illustrate decisions alongside their potential outcomes. Lastly, K-Nearest Neighbor (KNN) is a straightforward yet impactful instance-based learning method that classifies new instances by examining the predominant class among their k-nearest neighbors within the feature space. Together, these methods are essential tools in the domain of machine learning and artificial intelligence, enabling computers to scrutinize and comprehend intricate datasets. patterns in remarkably innovative ways, to classify text into abuse comments and non-abuse comments categories. Training these models involved incorporating a variety of features such as profane language, negative emotions, positive emotions, hyperlinks, proper nouns, and pronouns. However, they did not address the sub-types of abuse comments. Similarly, the researchers in [15] crafted an ensemble model for abuse comments identification. They incorporated The intricate and sophisticated frameworks known as Long Short-Term Memory (LSTM) networks, which are specifically designed to excel in the handling of sequential data and temporal dependencies, alongside the convolutional neural networks (CNN) that are adept in recognizing patterns and features within visual data, represent a powerful combination of advanced machine learning architectures that have vastly expanded the capabilities of artificial intelligence in various applications, ranging from natural language processing to image recognition and beyond, which have proven effective in spotting instances of abuse comments. Their findings affirm the method's efficiency in recognizing and categorizing offensive language across social media platforms. The authors acknowledge that there remains significant work to enhance the reliability of methods for detecting abuse comments, particularly the challenge of ensuring efficacy across various contexts. They advocate for further exploration of advanced methodologies and innovative technological applications to bolster abuse comments detection, highlighting the imperative for continuous advancements.

In [16], the researchers implemented three deep learning and six traditional algorithms for the classification of abuse comments. Their results indicated that LSTM emerged as the most accurate method for detecting abuse comments in terms of accuracy and recall. However, they did not address

issues related to class imbalance or granular classification of abuse comments types. Additionally, the research in [17] established a framework for identifying abuse comments within texts, utilizing a Fuzzy Logic System that leverages the outputs from SVM classifiers as inputs to pinpoint instances of abuse comments. Results indicated a need to enhance the precision of SVM classifiers to enhance the evaluation of bullying intensity via Fuzzy Logic. Furthermore, the limitation of this study lies in the difficulty of gauging the severity of bullying cases based on the analyzed tweets. Despite implementing a fuzzy logic system, the authors faced challenges in consistently determining the severity of bullying incidents. They noted that the subjective nature of assessing the severity of bullying episodes varied among authors, even when applying identical criteria to formulate fuzzy rules. This variability suggests that the authors did not always reach consensus on the seriousness of a bullying incident, rendering it a subjective and complex dimension of their research.

METHODOLOGY

To achieve accurate fine-grained classifications for abusive comment types, the model begins by employing a Multi-Layer Perceptron (MLP) classifier utilizing the One-Against-One strategy. This method enables the identification of complex patterns within the data related to abusive comments. Following this, class probabilities for each type of abusive comment are extracted, providing valuable insights into the likelihood of various forms of abuse. These probabilities are then processed through a Selection Set Algorithm, which effectively addresses the uncertainties and vagueness inherent in the classification task.

The Selection Set Algorithm enhances the model's capacity to navigate the ambiguity and overlapping characteristics of different categories of abusive comments. By leveraging this algorithm, the model establishes a more robust framework for classifying ambiguous content. Finally, the model synthesizes the results to make informed classification decisions, thereby refining the detection of abusive comments while accounting for the complexities of online communication. Figure 1 illustrates the key components of the model and their interconnections.

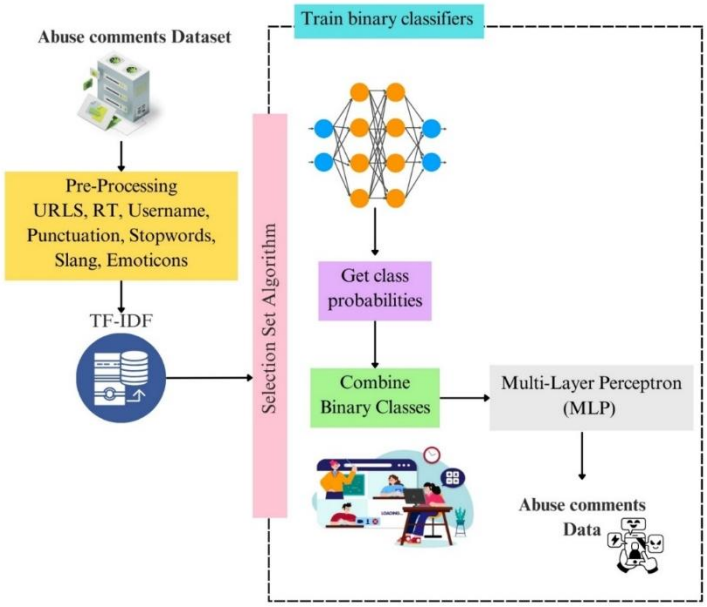

Figure 1. The Proposed Neutrosophic Abuse Comments Fine-Grained Classification.

Data Collection Phase

The data collection phase involves gathering a diverse set of comments from various social media platforms to ensure a comprehensive dataset. This includes scraping public posts, comments, and replies from platforms like Twitter, Facebook, and Instagram, focusing on both abusive and non-abusive comments. Care is taken to maintain ethical standards and comply with privacy regulations, ensuring that no personal information is collected without consent. The collected dataset is labeled according to the types of abuse, such as hate speech, harassment, and trolling, forming a foundation for subsequent analysis.

Table 1: Abusive comment dataset sample.

Comment ID	Comment	Type of Abuse
1	"You're such a loser, nobody likes you!"	Harassment
2	"Go back to where you came from!"	Hate Speech
3	"You're just too dumb to understand this."	Insult
4	"Shut up, no one cares about your opinion!"	Harassment
5	"This is the worst thing I've ever seen."	General Abuse
6	"Kill yourself, you're worthless!"	Severe Abuse
7	"Why are you so ugly?"	Insult
8	"People like you deserve to be bullied."	Hate Speech
9	"I hope you get what you deserve."	Threatening Language
10	"You are such a failure in life!"	Insult

Pre-Processing Phase

In the pre-processing phase, the raw comments undergo several transformation steps to prepare them for analysis. This includes text normalization, which entails converting all text to lowercase, removing punctuation, and eliminating stop words. Tokenization is performed to break comments into individual words.Moreover, methods like stemming or lemmatization are utilized to transform words into their fundamental roots. This phase also involves identifying and handling missing data or duplicates, ensuring the dataset is clean and representative of the varied expressions found in online interactions.

Granular Category Recognition Phase

The Granular Category Recognition Phase utilizes a Multi-Layer Perceptron (MLP) model with a One-Against-One strategy to distinguish between different types of abusive comments. Class probabilities are generated for each category, which are then processed using a Selection Set Algorithm. This algorithm effectively handles ambiguity, allowing for nuanced classifications that account for overlapping characteristics of abusive comments. The final output provides a detailed understanding of the types and likelihood of abusive behavior present in the dataset.

Binary Classification

Binary classification is all about discovering a function that can sort input vectors into one of two separate categories. Considering a training dataset Z-{(x_i,y_i):i∈1,…,l}, where every point x_i∈R^d embodies a characteristic vector along with the category identifiers y_i∈{-1,+1} signal the opposing and favorable categories, while the objective is to successfully distinguish between these two factions. The favorable category is represented by Z^+-{(x_i,y_i)∈Z┤ : ├ y_i-+1}, and the opposing category through Z^--{(x_i,y_i)∈Z:y_i--1}.

The goal is to employ a Multi-Layer Perceptron (MLP) to master intricate decision boundaries that separate the two classes. This boundary is characterized by a normal vector. w∈R^d and a bias b∈R. The MLP progressively fine-tunes the weights (w) and biases (b) through a process of minimizing classification errors throughout training, thus discovering the ideal parameters to differentiate the various classes.

Multi-Class Classification

In many practical applications, classifiers are tasked with distinguishing between n classes,where$n \in$ N. Given a training dataset $Z = (x_i, y_i): i \in 1, \dots, l$where each feature vector $x_i \in R^d$is associated with a class label, $y_i \in 1,2, \dots, n$, the primary objective of this endeavor is to meticulously engineer a function that possesses the capability to accurately categorize a given input vector into one of the predetermined n distinct classes. In light of the fact that the quantity of classes continues to expand, it becomes increasingly apparent that the intricacies associated with adapting binary classifiers to accommodate multiclass scenarios, particularly within the framework of large margin classifiers, experience a significant escalation in complexity.

Although it is widely recognized that binary classification presents a more straightforward and thus more efficient approach to data categorization, the endeavor of managing multiclass classification utilizing the identical dataset can introduce a level of complexity that is significantly greater and more intricate. Consequently, instead of attempting to directly modify and adapt binary classifiers for tasks involving multiple classes, it is frequently more beneficial and effective to deconstruct the multiclass dilemma into a series of distinct binary classification tasks that can be tackled individually. The subsequent outcomes derived from these individual binary classifiers are then meticulously aggregated and synthesized to ascertain the ultimate class label, which is essential for accurately categorizing the data. In this particular section, it shall delve into and elucidate two prominent strategies that have gained widespread acceptance and utilization in the realm of multiclass classification: the one-vs-one approach and the one-vs-all methodology, both of which offer unique advantages and considerations.

Selection Set Algorithm For One-Against-All (OAA) And One-Against-One (OAO) Classification

The Selection Set Algorithm is a robust classification approach designed to manage uncertainty and ambiguity in multi-class classification tasks, particularly in One-Against-All (OAA) and One-Against-One (OAO) frameworks. It enhances classification accuracy by leveraging the outputs of multiple classifiers and intelligently selecting the most relevant ones based on contextual information. In the OAA method, a separate binary classifier is trained for each class to distinguish it from all others. When a new instance is presented, each classifier generates a score indicating the likelihood that the instance belongs to its respective class. The scores are gathered into a vector S=[S1,S2,…,SN], where S_i is the score from classifier i. The Selection Set Algorithm applies a threshold T to filter these scores, retaining only those that exceed the threshold: $S' = \{S_i \mid S_i > T\}$. If no scores remain, the instance is classified as "unknown"; otherwise, the class corresponding to the highest score in S' is selected as the final classification $C_{\text{final}} = \arg\max_{C_i \in S'} S_i$. This approach effectively narrows down the potential classes, reducing noise from less relevant classifiers. In contrast, the OAO method involves training classifiers for every possible pair of classes. Each classifier votes for one of the two classes it has been trained to distinguish. During the prediction phase, votes from all classifiers are collected, resulting in a tally for each class. The Selection Set Algorithm enhances this voting mechanism by assigning weights based on the confidence levels of each classifier's prediction. A confidence threshold T_v is established to determine whether a vote is included in the final tally; votes from classifiers with confidence levels below this threshold are discarded. The total weighted votes for each class are computed, and the final classification is determined by selecting the class with the highest vote count, expressed as $C_{\text{final}} = \arg\max_{C_i} V_i$, where V_i represents the accumulated votes for class C_i. By integrating the Selection Set Algorithm into both OAA and OAO frameworks, the classification process is not only refined but also adapted to better handle ambiguous data. This adaptability is crucial in real-world scenarios, such as detecting abusive comments on social media, where the expressions may often overlap between categories.

Selection Set Algorithm
Start
Step 1: Input
Gather the following inputs:
Dataset with labeled instances.
Classifiers for either One-Against-All (OAA) or One-Against-One (OAO) classification.
Threshold values T (for OAA) and T_v (for OAO).
Step 2: Training Phase
- For One-Against-All (OAA):
Train a binary classifier f_i for each class C_i against all other classes.
- For One-Against-One (OAO):
Train a binary classifier f_{ij} for every pair of classes(C_i, C_j).
Step 3: Prediction Phase
For a new instance x :
ln OAA:
Compute scores S_i from each classifier f_i.
Create the score vector:

$$S = [S_1, S_2, \ldots, S_N]$$

In OAO:
Compute votes from each classifier f_{ij} for instance x.
Step 4: Selection Process
For OAA:
Filter scores using the threshold T :

$$S' = \{S_i \mid S_i > T\}$$

Check if S' is empty:
If empty, classify as "unknown."
If not, select the class with the highest score:

$$C_{\text{final}} = \arg\max_{C_i \in S'} S_i$$

For OAO:
Aggregate votes for each class.
Apply the confidence threshold T_v :
Include only votes from classifiers with confidence $> T_v$.
Determine the final classification:

$$C_{\text{final}} = \arg\max_{C_i} V_i$$

where V_i is the total votes for class C_i.

Step 5: Output
Return the final classification result C_{final} .
End

EXPERIMENTAL RESULTS

In this segment, the efficacy of the suggested framework is assessed ontwo datasets focused on abusive comments classification in social media, an arena where its prevalence and impact have grown considerably. The experiment was conducted using an The Intel (R) Core (TM) i3 CPU, equipped with 8.00 GB of RAM and integrated within Anaconda, serves as thisplatform. In this context, it adopt the assessment criteria outlined in [6]: Precision, Recall, and F1 Score for thisevaluation metrics [19].

$$\text{Precision} - \frac{T_P}{T_P + F_P} \tag{2}$$

$$\text{Recall} - \frac{T_P}{T_P + F_N} \tag{3}$$

$$\text{F1 Score} - 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Dataset	Total Tweets	Categories
Abusive Comments Dataset 1	47,000	Age, Ethnicity, Gender, Religion, Other Types of Abusive Comments, Not Classified as Abusive Comments
Abusive Comments Dataset 2	1,00,000	Race/Ethnicity, Gender/Sexual, Religion, Other Types of Abusive Comments, Not Abusive Comments

Performance Analysis for Fine Grained Abusive Comment Classification

Table 1 showcases the assessment outcomes of detailed abusive comment categorization across two distinct datasets, illuminating essential performance indicators: Precision, Recall, and F1 Score for multiple classifications. In Dataset 1, the model exhibits remarkable efficacy, especially within the age and ethnicity segments, attaining precision and recall figures of 0.98 and 0.99 correspondingly. The gender and Other_Abuse Commentsclasses show slightly lower metrics, with F1 scores of 0.91 and 0.87, indicating room for improvement. The overall accuracy for Dataset 1 is commendably high at 0.95. In Dataset 2, performance remains robust, especially for the ethnicity/race category, which achieves a perfect F1 score of 0.99. The gender/sexual and religion classes also perform well, boasting F1 scores of 0.98 and 0.91 correspondingly. The precision for Dataset 2 soars even higher at 0.97, highlighting the model's prowess in categorizing harmful remarks across various classifications given in table 2. 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, it were able to capture intricate relationships and nuances between different abuse comments 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.

Table 2. Evaluation Results Of Fine-Grained Abusive Comment Classification.

Data	Class	Precision	Recall	F1
	age	0.98	0.99	0.98
Dataset 1	ethnicity	0.99	0.98	0.98
	gender	0.89	0.92	0.91
	Other_abuse comments	0.88	0.87	0.87
	religion	0.99	0.97	0.98

	Accuracy			0.95
	ethnicity/race	0.99	0.99	0.99
Dataset 2	gender/sexual	0.99	0.98	0.98
	religion	0.89	0.92	0.91
	Accuracy			0.97

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 abuse comments classification tasks. By setting appropriate thresholds for truth, indeterminacy, and falsity memberships, it 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 abuse comments classification. The one-vs-one strategy provided a robust framework for capturing fine-grained distinctions between different abuse comments 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 abuse comments classification. By leveraging the strengths of both methodologies, our model demonstrated a superior ability to handle uncertainty, ambiguity, and overlapping features inherent in abuse comments data, resulting in precise and reliable classification results.

Comparative Analysis Of The Proposed Model And Existing Machine Learning Techniques

This series of trials was conducted to evaluate the effectiveness of the suggested model alongside various machine learning algorithms in the realm of nuanced abuse comment classification, utilizing a mix of certain machine learning techniques. The algorithms in question include Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR).

Table 3. Comparison Results of Different Machine Learning Methods On The Abuse Comment Dataset.

Algorithm	Class	Precision	Recall	F1
	Age	0.96	0.98	0.97
	Ethnicity	0.98	0.97	0.98
RF	Gender	0.92	0.84	0.88
	Other_Abuse Comments	0.78	0.90	0.84
	Religion	0.98	0.93	0.96
	Accuracy			0.92
LR	Age	0.99	0.96	0.97
	Ethnicity	0.95	0.97	0.96
	Gender	0.97	0.74	0.84
	Other_Abuse Comments	0.71	0.94	0.81
	Religion	096	0.90	0.93
	Accuracy			0.90
SVM	Age	0.98	0.98	0.98
	Ethnicity	0.98	0.98	0.98
	Gender	0.83	0.81	0.82
	Other_Abuse Comments	0.77	0.82	0.79
	Religion	0.99	0.96	0.97
	Accuracy			0.91

The selection of these algorithms stemmed from their widespread use and proven success in diverse classification challenges. Each algorithm exhibits unique advantages and drawbacks, with the aim being to gauge the performance of the proposed neutrosophic model in this context. The findings showcased in Table 3 validated that the suggested neutrosophic model surpassed the other machine learning algorithms regarding classification precision. The recommended combination led to an improvement in the accurate categorization of types of abusive comments.

Comparison Between The Proposed Model And Selection Set Using Abuse Comments Dataset

Table 4 encapsulates the assessment findings of nuanced abusive comment classification, following the deployment of the Selection Set Algorithm. This innovative algorithm significantly boosts the model's prowess in distinguishing between diverse categories of abusive comments, as highlighted by the showcased metrics: Precision, Recall, and F1 Score.

The outcomes derived from the classification process vividly demonstrate that the model exhibits exceptional performance across a diverse array of categories, showcasing its versatility and effectiveness in handling complex data. Within the Age segment, the model attains a remarkable precision rate of 0.98, coupled with a commendable recall of 0.97, which results in an outstanding F1 Score of 0.98, thereby reflecting its proficiency in accurately processing age-related information. In a similar vein, the Ethnicity segment reveals equally robust results, achieving an impressive precision of 0.99 alongside a recall of 0.97, which collectively culminates in an F1 Score of 0.98, thus underscoring the model's adeptness in precisely identifying and categorizing abusive comments pertaining to both age and ethnicity. These impressive statistics not only highlight the model's capabilities but also emphasize its critical role in fostering a more nuanced understanding of the factors influencing abusive language, reinforcing its significance in contemporary discourse analysis.

In the segment pertaining to Gender, there is a slight diminution in the precision metric, which has been recorded at an approximate value of 0.90, while the recall metric is noted to be around 0.87, leading to the overall calculation of an F1 Score that stands at 0.89. In contrast, the segment designated as Other_Abuse Comments reveals the most significant potential for enhancement, as evidenced by a precision score of 0.82 coupled with a recall score of 0.89, culminating in an F1 Score that is calculated to be 0.85. Thus, despite the areas identified for improvement, the robustness of the model in detecting various forms of abuse remains a noteworthy aspect of its overall efficacy.

Finally, the Religion category showcases remarkable performance, achieving a precision of 0.99 and a recall of 0.96, contributing to an F1 Score of 0.98. Collectively, the model attains an accuracy of 0.93, emphasizing the efficacy of the Selection Set Algorithm in honing the classification of abusive comments across distinct categories. This advancement reflects the algorithm's capability to handle uncertainty and ambiguity, ultimately bolstering the model's dependability in real-world scenarios.

Table 4. Evaluation Results Of Fine-Grained Abuse Comments Classification After Using Selection Set Algorithm

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_Abuse Comments	0.82	0.89	0.85
Religion	0.99	0.96	0.98
Accuracy			0.93

CONCLUSION AND FUTURE WORK

This paper suggested an accurate model for fine-grained abuse comments 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.

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. In summary, the proposed model for fine-grained abusive comment classification demonstrates superior performance compared to traditional in the vast and intricate realm of artificial intelligence, there exist an array of sophisticated machine learning algorithms, including but not limited to the highly esteemed Support Vector Machine (SVM), which excels at classification tasks by finding the optimal hyperplane for separating different classes; the

versatile Random Forest (RF), renowned for its ability to enhance predictive performance through the aggregation of multiple decision trees to reduce overfitting; and the widely utilized Logistic Regression (LR), which, despite its name suggesting a mere regression approach, actually serves as a powerful tool for binary classification by estimating probabilities using a logistic function, showcasing the diverse methodologies employed within the field to tackle a multitude of complex data-driven challenges. The integration of the Selection Set Algorithm with a Multi-Layer Perceptron (MLP) classifier significantly enhances classification accuracy, achieving an overall accuracy of 0.93 across various categories. Notable performance was observed in the Age and Ethnicity classes, both reaching F1 scores of 0.98, while the model effectively identified other forms of abuse, albeit with slight reductions in precision and recall. The ability to convert classification outputs into neutrosophic the establishment of various sets facilitates the opportunity for a significantly more intricate and multifaceted depiction of complex ideas and concepts that often require a deeper level of understanding and interpretation. uncertainty, which is critical in handling the complexities inherent in abusive comment classification. Overall, the results confirm that the combination of advanced machine learning techniques with neutrosophic logic principles leads to more reliable and precise classifications, underscoring the model's effectiveness in real-world applications related to online abuse detection

REFERENCES

[1] T. Ahmed, S. Ivan, M. Kabir, H. Mahmud, and K. Hasan, "Performance analysis of transformer-based architectures and their ensembles to detect trait-based abuse comments," *Social Network Analysis and Mining*, vol. 12, no. 1, p. 99, 2022, doi: 10.1007/s13278-022-00830-0.

[2] M. Alotaibi, B. Alotaibi, and A. Razaque, "A multichannel deep learning framework for abuse comments detection on social media," *Electronics*, vol. 10, no. 21, p. 2664, 2021, doi: 10.3390/electronics10212664.

[3] M. Arif, "A systematic review of machine learning algorithms in abuse comments detection: Future directions and challenges," *Journal of Information Security and Cybercrimes Research*, vol. 4, no. 1, pp. 1–26, 2021, doi: 10.18844/jiscr.v4i1.5583.

[4] R. Bayari and A. Bensefia, "Text mining techniques for abuse comments detection: State of the art," *Advances in Science, Technology and Engineering Systems Journal*, vol. 6, no. 1, pp. 783–790, 2021, doi: 10.25046/aj050949.

[5] A. Chhabra and D. K. Vishwakarma, "A literature survey on multimodal and multilingual automatic hate speech identification," *Multimedia Systems*, vol. 29, no. 3, pp. 1203–1230, 2023, doi: 10.1007/s00530-022-00892-1.

[6] R. K. Chilukuri, H. K. Kakarla, and K. S. Rao, "Radar signal recognition based on multilayer perceptron neural network," *International Journal of Electrical and Computer Engineering Systems*, vol. 14, no. 1, pp. 29–36, 2023, doi: 10.11591/ijece.v14i1.8962.

[7] T. D. Cox and J. Raditch, "Teaching online and abuse comments: Exploring abuse comments policies," *Journal of Effective Teaching in Higher Education*, vol. 5, no. 1, pp. 71–89, 2022, doi: 10.36021/jethe.v5n1.40.

[8] Y. Guo and A. Şengür, "A novel image segmentation algorithm based on neutrosophic filtering and level set," *Neutrosophic Sets and Systems*, vol. 1, no. 49, pp. 46–49, 2013, doi: 10.5281/zenodo.1166145.

[9] M. I. Islam, F. M. Yunus, E. Kabir, and R. Khanam, "Evaluating risk and protective factors for suicidality and self-harm in Australian adolescents with traditional bullying and abuse comments victimizations," *American Journal of Health Promotion*, vol. 36, no. 1, pp. 73–83, 2022, doi: 10.1177/08901171211017254.

[10] A. K. Jain, S. R. Sahoo, and J. Kaubiyal, "Online social networks security and privacy: Comprehensive review and analysis," *Complex & Intelligent Systems*, vol. 7, no. 5, pp. 2157–2177, 2021, doi: 10.1007/s40747-021-00266-5.

[11] B.-B. Jia, J.-Y. Liu, J.-Y. Hang, and M.-L. Zhang, "Learning label-specific features for decomposition-based multi-class classification," *Frontiers of Computer Science*, vol. 17, no. 6, p. 176348, 2023, doi: 10.1007/s11704-022-00283-x.

[12] C. S. Manigandaa, V. D. Ambeth Kumar, G. Ragunath, R. Venkatesan, and N. Senthil Kumar, "De-noising and segmentation of medical images using neutrophilic sets," *Full Length Article*, vol. 11, no. 2, pp. 111–111, 2023, doi: 10.11648/j.ijmr.20231102.12.

[13] E. W. Pamungkas, V. Basile, and V. Patti, "A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection," *Information Processing & Management*, vol. 58,

no. 4, p. 102544, 2021, doi: 10.1016/j.ipm.2021.102544.

[14] J. Pyżalski, P. Plichta, A. Szuster, and J. Barlińska, "Abuse comments characteristics and prevention—What can we learn from narratives provided by adolescents and their teachers?," *International Journal of Environmental Research and Public Health*, vol. 19, no. 18, p. 11589, 2022, doi: 10.3390/ijerph191811589.

[15] C. Raj, A. Agarwal, G. Bharathy, B. Narayan, and M. Prasad, "Abuse comments detection: Hybrid models based on machine learning and natural language processing techniques," *Electronics*, vol. 10, no. 22, p. 2810, 2021, doi: 10.3390/electronics10222810.

[16] J. Rao, L. Peng, J. Rao, and X. Cao, "Modified taxonomy method for double-valued neutrosophic number MADM and applications to physical education teaching quality evaluation in colleges and universities," *Journal of Intelligent & Fuzzy Systems*, vol. 44, no. 6, pp. 10581–10590, 2023, doi: 10.3233/JIFS-211314.

[17] T. Wullach, A. Adler, and E. Minkov, "Towards hate speech detection at large via deep generative modeling," *IEEE Internet Computing*, vol. 25, no. 2, pp. 48–57, 2020, doi: 10.1109/MIC.2020.2986588.