IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 July 2024, accepted 7 August 2024, date of publication 16 August 2024, date of current version 11 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3445288

RESEARCH ARTICLE

Cyberbullying Detection and Abuser Profile Identification on Social Media for Roman Urdu

AYESHA ATIF[1], AMNA ZAFAR[1], MUHAMMAD WASIM[2], TALHA WAHEED[1], AMJAD ALI[3], HAZRAT ALI[4], (Senior Member, IEEE), AND ZUBAIR SHAH[3]

[1]Department of Computer Science, University of Engineering and Technology Lahore, Lahore 39161, Pakistan
[2]University of Management and Technology, Sialkot Campus, Lahore 51310, Pakistan
[3]Division of Information and Computing Technology, College of Science and Engineering (CSE), Hamad Bin Khalifa University (HBKU), Qatar Foundation, Doha, Qatar
[4]Computing Science and Mathematics, University of Stirling, FK9 4LA Stirling, U.K.

Corresponding authors: Amjad Ali (amsali@hbku.edu.qa) and Zubair Shah (zshah@hbku.edu.qa)

Open Access funding provided by the Qatar National Library. This work was supported by the Higher Education Commission (HEC), Pakistan, through the National Research Program for Universities (NRPU) under Grant 20-16597/NRPU/R&D/HEC/2021.

ABSTRACT In today's digital era, the escalating phenomenon of cyberbullying is a pervasive and growing concern. With the increasing prevalence of social media platforms, such as Twitter, online abusive behavior has become a significant issue that often leads to unpleasant experiences for users. Manual detection of abnormal and bullying behavior within the realm of social media is inherently not scalable. Moreover, most existing studies on cyberbullying detection have been predominantly conducted in English and very limited work has been done on Urdu (a widely used language in Asia). This paper presents an approach for detecting cyberbullying in Roman Urdu tweets and identifying abuser profiles on Twitter. Firstly, we develop a text corpus of Roman Urdu tweets with user profile data. Subsequently, we employ Gated Recurrent Unit (GRU) model coupled with the application of word2vec technique for word embedding to develop a cyberbullying detection model. Furthermore, we present temporal abusive tweet probability analysis method to provide a nuanced analysis of the number of bullying and non-bullying tweets sent by individuals within a specific time interval. To evaluate the performance, we compare the GRU-based approach with other machine learning models. The results show that the GRU model with lexical normalization gives the best results with an accuracy of 97% and F1-measure of 97%.

INDEX TERMS Cyberbullying detection, social media, Roman Urdu, machine learning, deep learning, abuser profile identification.

I. INTRODUCTION

With the advancement in technology and digital inventions, usage of social media is increasing day by day. Billions of users are actively engaging with social media platforms such as Twitter (now X) and Facebook to express their thoughts and ideas. According to a report, Twitter has 338 million active users [1], witnessing an average of 200 billion tweets annually. Around 68.1% of Twitter users are male, and 31.9% of Twitter users are female. These platforms serve as virtual arenas for discourse, allow users to share their thoughts on a wide spectrum of topics related to politics, religion, sports, events, and current affairs. Unfortunately, some people use these platforms for spreading hate and posting bullying contents. The cloak of anonymity provided by social platforms, wherein users commonly adopt pseudonyms instead of real names, has contributed to a notable surge in cyberbullying posts. Consequently, surveillance and monitoring of such posts is becoming more challenging.

Cyberbullying remains a critical issue on the Internet, with a concerning number of individuals, especially teenagers, falling victim to its deleterious effects. According to a survey by Security.org,[1] a staggering 44% of users experienced online harassment, especially teenagers and kids. Thus,

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

[1]https://www.security.org/resources/cyberbullying-facts-statistics/

cyberbullying has morphed into a significant problem that affects people's well-being and mental health, especially teenagers. Cyberbullying comes in many forms, such as the use of offensive language, harassment, racism, hate remarks about one's choice of lifestyle/religion, personal criticism, etc. While social media platforms may dedicate a support team that enables reporting any abusive content/accounts, expeditious response to such reporting is not always guaranteed, thus failing to prevent the damage or satisfactorily address the affected person's concerns. Therefore, it is highly recommended to develop methods to identify cyberbullying content. While quite a few deep learning and machine learning approaches have been developed for cyberbullying detection; these methods are primarily developed for the English language [2], [3], [4], [5], [6]. On the contrary, minimal work has been done in the Roman Urdu language, which is a widely used language in Asia and commonly used on the social media platform as well [7], [8]. To redress this gap, we present an approach to detect cyberbullying in Roman Urdu text on social media using deep learning. We make the following contribution in this paper:

- We develop a corpus of Roman Urdu tweets for cyberbullying detection. We call it the RU dataset. The dataset is preprocessed using the pre-trained word embedding technique FastText, trained over 157 languages. We also compare its performance with other methods, such as Term Frequency Inverse Document Frequency (TF-IDF) and word2vec.
- We develop a framework using the GRU model with word2vec embeddings for detecting cyberbullying in Roman Urdu tweets and identifying abuser profiles on Twitter using a computational method.
- We evaluate the performance of lexical normalization of words and compare the performance of the GRU model with other machine learning methods for the cyberbullying detection task. Our approach using the GRU model with lexical normalization achieves the best performance with an accuracy of 97% and f1-measure of 97%. Finally, we use simple heuristic to identify abusive profiles, thus, categorizing the users into normal, suspected, and abusive users.

The rest of the paper is organized as follows: Section II presents a literature review of the existing methods for cyberbullying detection and classification, focusing on different languages along with abuser profile identification. Section III explains the RU dataset development and annotation. Section IV explains the proposed framework in detail, including the preprocessing of the data, the GRU-based pipeline, and the GRU model training. Section V presents the experiments and comparisons with other machine learning models and discusses the results. Finally, Section VI concludes the paper with suggested future work.

II. RELATED WORK

Cyberbullying and the use of offensive language on social media, particularly on Twitter, has become a major concern. Various studies have been conducted to develop methods for detecting these issues. In this section, we discuss the existing state-of-the-art research undertaken in cyberbullying and offensive language detection in multiple languages. We have categorized the existing research work into three categories: 1) machine learning-based approaches, II) deep learning-based approaches, III) methodologies dedicated to abuser profile identification.

A. MACHINE LEARNING APPROACHES

In the past, simplistic lexical methods were once deemed efficacious for identifying offensive and bullying language. However, recent research has demonstrated that, aside from specific terms, the efficacy of these lexical-based methods is limited. In comparison, machine learning algorithms exhibit superior speed and formidable performance compared to these rudimentary lexical normalization approaches. Conventional machine learning models tend to excel in scenarios with small datasets, yet their performance degrades when dealing with large-scale datasets. Notably, in the work by Rasheed et al. [9], a Linear Support Vector Machine (SVM) was proposed for hate speech detection in Roman Urdu tweets, surpassing the performance of other models, such as those utilizing TF-IDF and Count-Vectorizer feature extraction techniques, by a considerable margin. It's important to note, however, that this study was confined to binary classification. Another SVM-based investigation, as detailed in Alduailaj and Belghith [10], examined cyberbullying detection using the Arabic language in conjunction with TF-IDF feature extraction. The authors introduced a novel cyberbullying detection technique named ''Passive Regressor''. This innovative approach not only had the advantage of improved interpretability but also demonstrated robust performance in cyberbullying detection specifically tailored to the Bengali language. Similarly, another investigation highlighted that leveraging Count-Vectorizer features in conjunction with logistic regression yielded an impressive overall F1-score of 90% for hate speech detection [14]. This study employed machine learning techniques for identifying hate speech in Roman Urdu, and the findings indicate that logistic regression exhibits superior performance in distinguishing between hostile and neutral tweets. Furthermore, in the research by Sreelakshmi et al. [15], a Support Vector Machine (SVM) employing the Radial Basis Function (RBF) kernel, combined with character-level FastText, was proposed for the purpose of detecting hate speech in code-mixed social media texts that include both Hindi and English. The study revealed that utilizing character-level features from FastText provided more information for classification compared to the traditional word and document-level features.

B. DEEP LEARNING APPROACHES

Deep learning models have demonstrated superior performance, particularly when dealing with large datasets, and they also excel in handling imbalanced data. In the realm of abusive content detection on Twitter, a hybrid model

TABLE 1. Summary of previous works.

Ref.	Methodology	Features	Dataset	Targeted Categories	Language	Evaluation Score	Contribution
[9]	SVM	TF-IDF, Count-Vectorizer	50000 tweets	bully, non-bully	Roman Urdu	acc:97.8%	1-Dataset of Roman Urdu Tweets. 2-Performance evaluation of baseline ML classifiers with different feature extraction techniques.
[10]	SVM	TF-IDF, BoW	30,000 tweets	Bully, non-bully	Arabic	acc:95%	1- A dataset of Arabic text. 2- Performance evaluation of the SVM ML classifier on the Arabic tweets.
[5]	Dolphin Echolocation Algorithm (DEA) and RNN	Word2Vec, TF-IDF	10,000 tweets	Cyberbully, non-cyberbully	English	prec:89.52% rec:88.98% f1:89.25% acc:90.45%	1- Hybrid approach using DEA and Elman RNN for fine parameters tuning. 2- Efficiency comparison with existing ML models and the Bi-LSTM model.
[6]	SVM, NB, DT, MLP, RF	Word2Vec, TF-IDF, FastTex	24783 tweets	Offensive, non-Offensive	English	prec:95% rec:94% f1:95% acc:93%	1- A text-classification pipeline using supervised learning approaches to present the best-performing on the Twitter dataset.
[11]	Bi-LSTM with attention layer	Word2Vec	30,000 tweets	Hate, neutral	Roman Urdu	acc:87.5%	1- A context-aware approach for Roman Urdu hate speech analysis using bi-directional LSTM with lexical normalization. 2- Performances analysis with another ML algorithms with and without lexical normalization.
[12]	Bi-LSTM with attention layer	Word2Vec, Glove, FastTex	26,824 user reviews	Positive, negative, neutral	Roman Urdu	acc: 67%	1- Proposed an RU-BiLSTM for roman Urdu sentiment analysis 2- Developed a larger roman Urdu e-commerce reviews-based dataset.
[4]	DNN with three dense layers	Word2Vec, emotional feature	Cyber-troll (20,000)	Aggressive, non-aggressive	English	prec: 86.28% rec:87% f1:87.11% acc:88%	1- Proposed a multi-layer model using Sigma and RELU activation functions for aggressive content detection. 2- Proposed word2Vec-based feature selection method.
[13]	SVM, PR, RF, LR	TF-IDF, Word2Vec	Not mentioned	Bully, non-bully	Bengali	Passive aggressive classifier: acc:78.1%	1- Proposed a text-classification method for regional Bengali language using ML.
[14]	Logistic regression	Word embedding, TF-IDF, N-Gram Vector	HS-RU-20 (5000 tweets)	Offensive, neutral	Roman Urdu	LR with CV perform best : acc: 81%	1- Contributed a public dataset of 5000 Roman Urdu hate speech tweets. 2- Performance analysis using ML and DL approaches for hate speech detection.
[15]	SVM-Radial basis function	FastText	10,000	Hate, non-hate	Hindi-English mixed data	acc: 85.81% prec:85.86% rec:85.81% f1:85.80%	1- Analysis showed that FastText character level features provide more code mixed test classification than words and document level features.
[16]	CNN-gram with four CNN layers	BERT, FastText, RomUrEm	RUHS-OLD (10,012 tweets)	Abusive, sexism, religious hate, profane, normal	Roman Urdu	acc:82% prec:75% rec:74% f1:75%	1-Developed annotated dataset of 10,000 Roman Urdu tweets. 2-Proposed CNN-gram for offensive language detection and hate speech.

Note. acc: Accuracy. CV: Cross Validation. prec: Precision. rec: Recall.

utilizing the Dolphin Echolocation Algorithm (DEA) in combination with a Recurrent Neural Network (RNN) featuring word2vec embeddings was proposed [5]. Some other research studies [5], [11] affirm the increased effectiveness of neural learning-based approaches in detecting hate speech when compared to classical machine learning techniques. Another study [11] focused on the identification of hate speech in Roman Urdu text. The study revealed that among various neural network architectures, the Bidirectional Long Short-Term Memory model (Bi-LSTM) coupled with the Adam optimizer, and further enhanced by lexical normalization, exhibited the most robust performance. This work emphasized the significance of lexical normalization in Roman Urdu due to the language's numerous variations,

indicating that using Bi-LSTM without lexical normalization could result in overfitting. In another study [12], the authors explored the utilization of Bi-LSTM architecture augmented with an attention layer for detecting negative user reviews. This research highlighted that employing Bi-LSTM with multiple embedding techniques within the neural network architecture led to notable improvements in hate speech or negative comment detection.

Khan et al. [4] used a combination of word embedding and eight different emotional features to feed into the DNN model and achieved f1-score of 97%, which surpassed that of various machine learning and deep learning models. This study also suggested that embedding emotional features in deep learning models can give better results for Roman Urdu. Rizwan et al. [16] proposed a deep learning-based architecture called ''CNN-gram'' for detecting hate speech and abusive language in Roman Urdu. The proposed model is evaluated and compared against several baseline machine-learning models on the RUHSOLD dataset. A hybrid model that combines the GRU layers and CNN layers for hate speech detection is proposed by Zhang et al. [19]. The authors suggested that GRU is the same as LSTM with a lesser layer which leads to better performance in less time. However, this technique is limited only to English language. Previous work has predominantly focused on detecting Offensive and hate speech for English language and other regional languages across the globe [20]. Some studies have utilized their native languages for this purpose, but there is limited research on detecting cyberbullying contents in Roman Urdu.

To best of our knowledge, no study has fully explored detection of cyberbullying using Roman Urdu tweets and identifying abuser profiles on Twitter. Moreover, available datasets and standard dictionaries for Roman Urdu cyberbullying text are also limited. Thus, in this paper, firstly, we develop a corpus of Roman Urdu tweets for cyberbullying detection called RU dataset. The dataset is preprocessed using the pre-trained word embedding technique FastText. Then, we develop a framework using the GRU model with word2vec embeddings for detecting cyberbullying using Roman Urdu tweets and identifying abuser profiles on Twitter using a computational method.

C. ABUSER PROFILE IDENTIFICATION

In addition to the detection of bullying and hate speech content on social media platforms, there is growing demand to identify the users' profiles that are involved in cyberbullying activities [21], [22]. Limited number of studies have focused on this area as summarized in Table 2. Nurrahmi and Nurjanah [17] proposed a probabilistic model to measure user credibility. The study suggested a model for the detection of abusive tweets in the Indonesian language [17]. Based on a threshold of abusive content posted in specific time intervals, the user is categorized as bullying or normal user. Sarna and Bhatia [23] adopted a distinct method that involved machine learning models and graphs to identify cyberbullying users. In this approach, the probability of cyberbullying contents in a user's message was utilized. Additionally, they suggested eight guidelines to extract essential features during the process of cyberbullying tweet classification, which could be improved for Roman Urdu.

In a study done on Arabic tweets, based on the type of tweets, the user's profile data was analyzed by Abozinadah et al. [18]. The study used some profile features and social networking graphs associated with bullying tweets to measure the degree of abusive content and profile identification [18]. While the studies mentioned in this section have made progress in identifying abusive users on social media platforms, there are still several limitations that need to be addressed. One limitation is the language-specific approach used in these studies, which may not apply to other languages. Moreover, no results validation strategy or guidelines are available in previous studies. To overcome these challenges, we propose a computational model to detect abusive profile posting bullying content in Roman Urdu and an evaluation strategy for the result's validation.

III. RU CYBERBULLYING DATASET

A. LIMITATIONS IN EXISTING DATASETS

The linguistic landscape of Roman Urdu is characterized by scarcity of resources, resulting in limited efforts towards developing annotated datasets containing cyberbullying content [11], [14], [16]. Furthermore, the availability of pre-trained embedding techniques for Roman Urdu remains constrained. To address these gaps, [11] proposed annotation guidelines to enable contextual analysis of Roman Urdu data and curated a substantial dataset of 30,000 tweets, annotated by domain experts. In a separate study focusing on hate speech detection [14], the authors adopted an iterative approach to formulate annotation guidelines for the HS-RU-20 dataset, tailored specifically for hate speech detection in Roman Urdu. Additionally, Rizwan et al. [16] explored the development of the Roman Urdu corpus and examined the performance of five distinct multi-lingual pre-trained embeddings, including FastText, LASER, BERT, ELMo, and XLM-RoBERT, on the Roman Urdu dataset. The study also proposed a pre-trained embedding technique named as RomUrEm, but it is not available for researchers to study and reuse.

B. DATASET DEVELOPMENT

As per the existing literature, the construction of bullying datasets typically involves extracting offensive content from online sources using lexicons containing abusive words or utilizing hate-speech datasets. Nevertheless, while Hatebase.org provides an extensive collection of multilingual bullying words, it lacks a lexicon base for the Roman Urdu language. Additionally, the absence of profile data associated with Roman Urdu tweets in existing datasets presents a further challenge. In response to this gap, we have formulated a lexicon of abusive words for Roman Urdu by conducting online keyword searches as shown in figure 1. This

TABLE 2. Summary of abuser profile identification work.

Ref.	Methodology	Dataset	Dataset Features	Language	Contribution
[17]	Probabilistic model for user credibility	5000 tweets	tweets content	Indonesian	1- Probabilistic model for user credibility using a directed graph.
[18]	Naive Baye's classifiers model and social networking feature	Not mentioned	tweet, profile information	Arabic	1- First dataset for Arabic abusive accounts with Arabic tweets. 2- A novel approach for abusive accounts detection using multiple features (tweet, profile content, social graph).

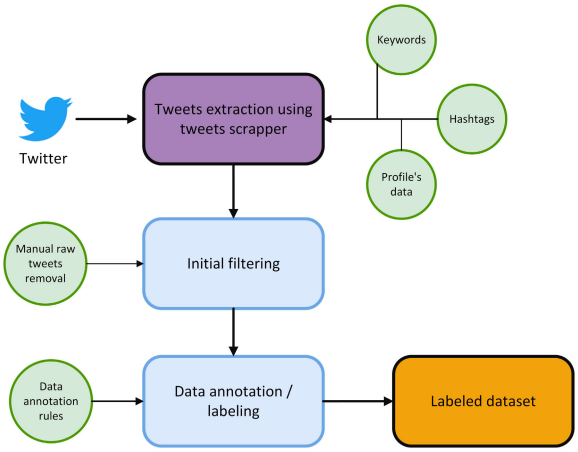

FIGURE 1. Dataset collection and labeling.

TABLE 3. Number of tweets in each category.

Category	No. of tweets
Neutral	5355
Racism	822
Abusive/Offensive	3241
Sexism	580
Religious Hate	672

comprehensive lexicon encompasses terms associated with bullying, religious hate, racism, and sexist language. After extracting a corpus of Roman Urdu tweets using lexicons and hashtags, labeling annotations were established. These annotations considered diverse categories of cyberbullying as discussed in various articles. To ensure unbiased results, class balancing is performed. The dataset contains an equal number of bullying and non-bullying tweets, as depicted in Table 3.

In our dataset of tweets labeled as either "bullying" or "non-bullying" in Roman Urdu, there's a recognition that labeling can be influenced by personal interpretation and cultural context. Different people may view the same tweet differently, leading to labeling inconsistencies. This happens because the perception of bullying varies based on personal experiences and cultural norms. Despite having guidelines for annotators, the subjective nature of bullying contents may still introduce errors in labeling. These errors can the generalizability of the machine learning models trained on the data, making it important to address them through strategies like ongoing quality checks and incorporating uncertainty measures in model predictions.

C. DATASET ANNOTATION

We annotated the dataset into five distinct categories: Neutral, Offensive, Religious Hate, Sexism, and Racism. A selection of tweets from the dataset, along with their corresponding labels, is shown in Table 4. These annotations were conducted with strict adherence to the guiding principles defining each target category.

- **Sexism:** This category encompasses any manifestation of online aggression or harassment directed towards an individual based on their gender or sex [21], [24].
- **Offensive/Abusive:** This category encompasses any usage of vulgar language or aggression with the intent to inflict harm or distress upon the victim [22], [24]. This may involve derogatory language or threats.
- **Racism:** This category entails any form of bullying or aggression directed towards an individual or a group based on their race or ethnicity [24].
- **Religious Hate:** This category includes any form of aggression driven by prejudice or discrimination against an individual's religious beliefs or faith system [25]. It involves propagating hate against a specific religious community or religious groups.
- **Neutral:** This category refers to tweets that do not fall into any of the aforementioned categories. It encompasses simple tweets that contain no harmful or hurtful language.

IV. METHODOLOGY

This section describes our proposed method in detail as illustrated in Figure 2. The proposed architecture encompasses the following six phases.

1) Roman Urdu tweets scrapping.
2) Dataset preprocessing.
3) GRU model training for classification.
4) GRU model evaluation.
5) Abuser profile identification.

A. ROMAN URDU TWEETS SCRAPPING

For scraping tweets, we utilized Apify.[2] The data selection was based on the top Twitter trends, encompassing subjects such as politics, harassment, events, showbiz, and sports. Data retrieval was accomplished through hashtags, keywords,

[2]Apify is a Twitter scraping tool available at https://apify.com/store/scrapers/twitter

TABLE 4. Sample of tweets from dataset with targeted label.

Tweets	Category/Label	Description
Tm pathan ho tumhen kya maloom k aqal kya hoti ha.	Racism	This is targeting a specific race i.e. Pathan.
bakwas bnd kr or tu b chali ja.	Abusive/Offensive	This tweet contains aggressive word.
Bilawal hijra ha daikh to bol kese rha ha.	Sexism	Tweet contains aggressive word targeting specific gender.
ra***i ki olaad. Tere jesi boohat ortain daikhi hai main ne.	Abusive/Offensive	Tweet contains offensive word.
Tum hindu ho he Har**i, tmhara koi aik baap nahi hota.	Religious Hate	Tweet is targeting specific ethnicity based on religious differences.
or kya ho rha ha.	Neutral	Tweets doesn't contain any form of aggression or offensive word.

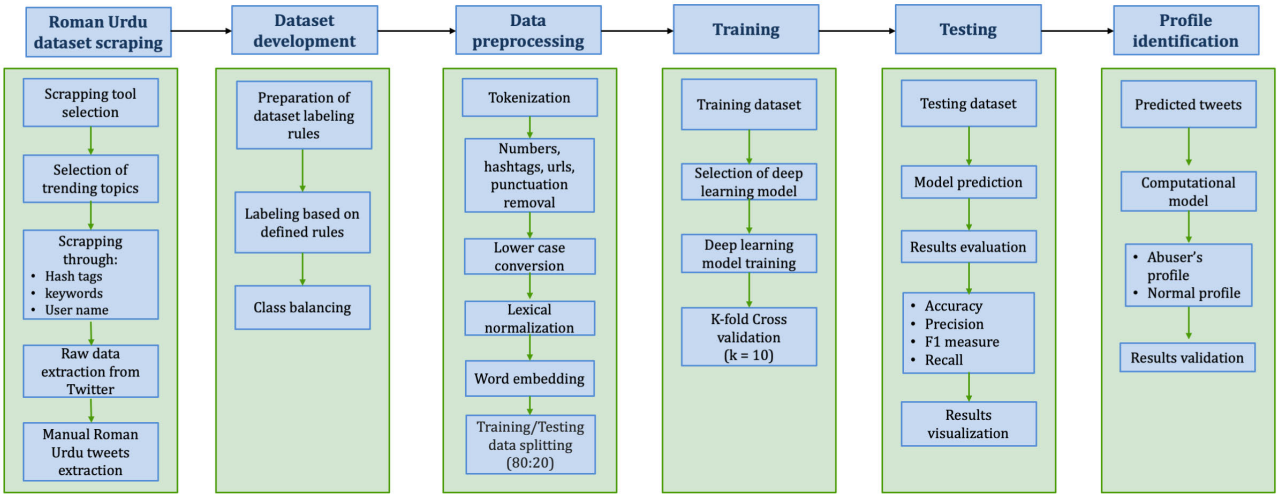

FIGURE 2. Proposed architecture of cyberbullying text detection and abuser profile identification.

and user profiles. The selection of keywords was determined by data availability within each category, which includes racism, sexism, religious hate, abusive content, and neutral tweets. Initially, a total of 36,000 raw tweets were extracted from Twitter. Subsequently, manual filtering was performed to include only Roman Urdu tweets, resulting in 10,670 tweets as illustrated in Table 3.

B. DATASET PREPROCESSING

Before feeding data into the deep learning model, we preprocessed the data to translate it into the required format. The preprocessing steps are explained below.

1) **Tokenization:** During this step, each tweet underwent tokenization into individual words using the TweetTokenizer API from the NLTK library.[3]
2) **Data Cleaning:** This step focused on cleaning the extracted Roman Urdu tweets. It was observed that tweets contained various special characters, URLs, mentions, punctuation marks, hashtags, numbers, and emojis expressing emotions. These elements were removed from the dataset using a Python code script.
3) **Lower Case Conversion:** To ensure standardized and case-insensitive text for the model, all letters were converted to lowercase.
4) **Stop Words Removal:** Stop words refer to irrelevant words within sentences, such as conjunctions and prepositions. As Roman Urdu data is already sparse, removing stop words helps to avoid unnecessary dimensionality in the dataset. We utilized Roman Urdu stop words available at[4] to filter them out from our dataset.
5) **Lexical Normalization:** Given that Roman Urdu lacks a standard writing style or set of words due to regional and individual variations, lexical normalization becomes crucial. These lexical variations result in sparse data and can affect the model's performance. For instance, the word **[world]** is written as "kainat," "kaenaat," "kayenat," "kaynat," or "kainaat" in Roman Urdu. This variation is known as lexical variation. Reducing lexical variation helps alleviate data sparsity and improves the model's performance. To ensure consistency and handle variations in Roman Urdu terms, lexical normalization is crucial. This process involves translating different spellings and forms of a word into a standardized lexical form, as depicted in Figure 4. For this purpose, we utilized a lexicon corpus of 61,000 Roman Urdu words for normalization of lexical variants, available at.[5]

[3]https://www.nltk.org/

[4]https://github.com/haseebelahi/roman-urdu-stopwords

[5]https://github.com/mtk12/Roman-Urdu-Lexical-variation-via-Clustering/

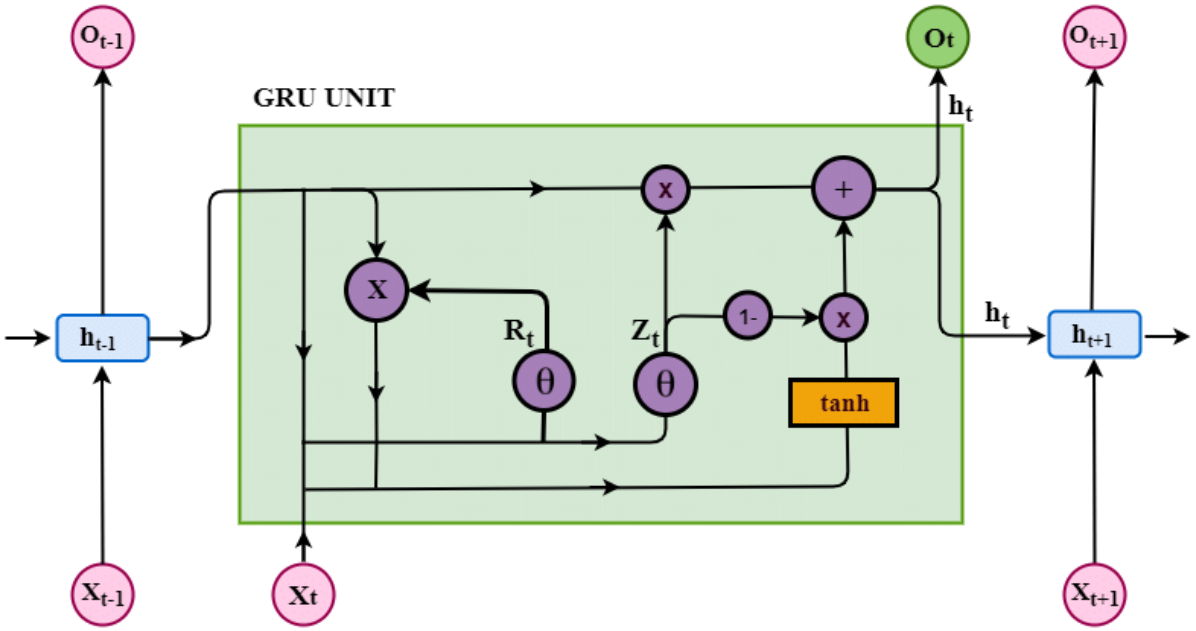

FIGURE 3. GRU architecture.

6) **Word Embedding using word2vec:** Following preprocessing and dataset normalization, the next vital step in raw data analysis involves feature extraction. Instead of manipulating raw data directly, the computer converts it into derived numerical values while preserving the original data's information. Various techniques can be employed for feature extraction, such as Bag of Words, TF/IDF with n-grams, and character gram. In this study, we employed the word2vec pretrained embedding approach utilized by the selected model, GRU, with a dimension size of 100 and a vocabulary size of 10,000. The dimension size was set to train the model over almost the same length of tweets. The word2vec technique transforms each word in the corpus based on its contextual meaning into a 100-dimensional vector. This process allows semantically related words to be grouped together in the vector space.
7) **Train/Test Data Splitting:** The data was split into an 80:20 ratio for training and evaluating the selected deep learning model.

C. PROPOSED GRU-BASED PIPELINE

Our proposed model is centered around the utilization of the GRU with a pre-trained embedding technique using word2vec. The GRU, a variant of Recurrent Neural Network (RNN), is specifically designed to tackle the vanishing gradient problem, commonly encountered in traditional RNNs. This problem arises when gradients in the network become exceedingly small during back-propagation, leading to slow training and the inability to learn long-term dependencies effectively. To address these challenges, the

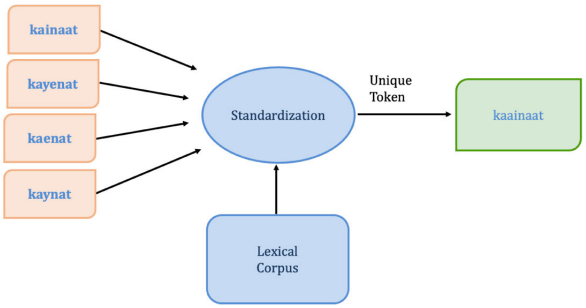

FIGURE 4. Dataset lexical normalization.

GRU incorporates gating mechanisms that facilitate selective updating and resetting of the hidden state, enabling it to capture long-term dependencies more efficiently in sequential data. Additionally, the GRU possesses a reduced parameter count compared to conventional RNNs, which in turn helps in preventing overfitting and improving the model's generalization performance. The fundamental GRU model consists of multiple cells and two gates: an update gate (z_t) and a reset gate (R_t), as illustrated in Figure 3. At each time step (t), the model takes input (x_t) and the previous hidden state (h_{t-1}) to compute the new hidden state (h_t) using the following equations for a single GRU cell:

$$h_t = tanh(Wx_t + R_t * Uh_{t-1}) \quad (1)$$
$$Z_t = \sigma(W_z x_t + U_z h_{t-1}) \quad (2)$$
$$R_t = sigma(W_r x_t + U_r h_{t-1}) \quad (3)$$

The update gate Z_t regulates the extent to which the candidate activation h_t influences the current state h_t, whereas the reset

gate r_t governs the extent to which the previous hidden state h_{t-1} is retained in the current state. The candidate activation h_t is a fusion of the input at time t and a modified version of the previous hidden state h_{t-1}, where the transformation is controlled by the reset gate R_t. The architecture of the GRU model utilized in our study is depicted in Figure 5. Tokenized tweets and padded sequences are first passed through an embedding layer with embedding dimensions of 200. This layer maps each word index to a dense vector (word embedding) and is initialized with pre-trained embeddings, forming an $\mathbf{l} \times \mathbf{d}$ dense embedding matrix, where $\mathbf{d}$ represents the embedding vector dimension for each word, and $\mathbf{l}$ is the number of words in the tweet. The word embeddings are then fed into the first GRU layer, which captures sequential information within the text. The output of the first GRU layer is subsequently passed to the second GRU layer for further processing, capturing more complex patterns in the data. The output of the second GRU layer is then fed through a dense layer with a softmax activation function, generating probability distributions over the class labels.

The model predicts the final class label by selecting the one with the highest probability from the output of the softmax layer. To optimize the model, we use categorical cross-entropy as the loss function, suitable for multi-class classification problems with probability distributions as outputs. For training, we employ the Adam optimizer, known for faster convergence and improved performance. To ensure consistent experimental conditions, all network weights are initialized randomly, and a fixed random seed is applied across all experiments. The model is trained for 10 epochs in each experiment, and a saved checkpoint of the learned weights is utilized to evaluate the test split based on the epoch with the best predictive performance on the validation split. If the validation error does not decrease for three consecutive epochs, the training process is terminated.

D. GRU MODEL TRAINING FOR CLASSIFICATION

In our work, we use GRU with pre-trained embedding to detect cyberbullying in Roman Urdu. The model was trained at this stage using a validation split of 80:20 ratio. By utilizing k-fold cross-validation, bias was significantly reduced as the majority of the data was employed for fitting. The results of training during k-fold cross-validation are recorded for purposes of visualization. In each epoch, the same data is given to the neural network multiple times to learn various features from the text. Data is distributed randomly so that data is not trained over the same pattern. During the data preprocessing step, the dataset was split into training and testing sets. A portion of the data was used for model training, while the remaining portion was reserved for model testing. The testing dataset underwent cleaning, lexical normalization, and word embedding processes. Subsequently, the GRU model predictions were generated for the testing dataset, and the model's performance was evaluated using metrics such as accuracy, precision, recall, and F1-score. The evaluation results were aggregated in tabular format to facilitate a comprehensive comparison with other baseline machine learning models.

E. PROFILE IDENTIFICATION

Following the training and testing of the model on the dataset, tweets were categorized into different types of cyberbullying activities. These categorized tweets were then employed to identify abuser profiles based on discernible patterns in the data. The identification of the abuser profile is primarily reliant on the frequency of categorized tweets and the language employed. Further details on the computational model for abuser profile identification are provided in Results section.

V. EXPERIMENTS AND RESULTS

In this section, we present the experiments and the corresponding results. In our work, we used traditional machine learning models using different embedding techniques and also compared the results with our proposed lexical normalization-based GRU deep learning model with two hidden layers. In the dataset prepration, out of the 10,670 tweets, 5,355 were labeled as ''Neutral'', while the remaining were categorized as bullying tweets. The bullying tweets were further divided into the following categories: ''Abusive/Offensive (3,241)'', ''Sexism (580)'', ''Religious Hate (672)'', and ''Racism (822)''. The dataset underwent tokenization, preprocessing, and normalization before being fed into the selected model. We employed a pre-trained word2vec embedding for feature extraction within our proposed pipeline. The deep learning models were implemented in Python using Scikit-learn, Keras, and Tensorflow libraries. Additionally, we employed Python packages such as JSON, Gensim, NLTK, Pandas, and Numpy.

A. TRADITIONAL MACHINE LEARNING MODELS

Initially, we employed six traditional machine learning classifiers as baseline models to construct multi-classifier cyberbullying detection systems. The selected machine learning models were Support Vector Machine (SVM), Multi-nomial Naive Bayes (NB), Decision Tree (DT), Logistic Regression (LR), Random Forest (RF), and Gradient Boosting (GB). These models were chosen based on a comprehensive literature review, indicating their effectiveness in hate speech or bullying data detection. We split the dataset into training and test sets with a ratio of 80:20. Each model was trained multiple times on the training set, utilizing a 10-fold cross-validation technique. We conducted the same experiment thrice for each model, employing FastText Embedding, TF/IDF, and word2vec with a vector size of 200 dimensions. The results for each scenario are summarized in Table 5. Figure 7 visually presents the performance of each machine learning model.

B. GRU MODEL PERFORMANCE

Figure 6 depicts the training and validation accuracy of the GRU Model, which progressively increases with each epoch

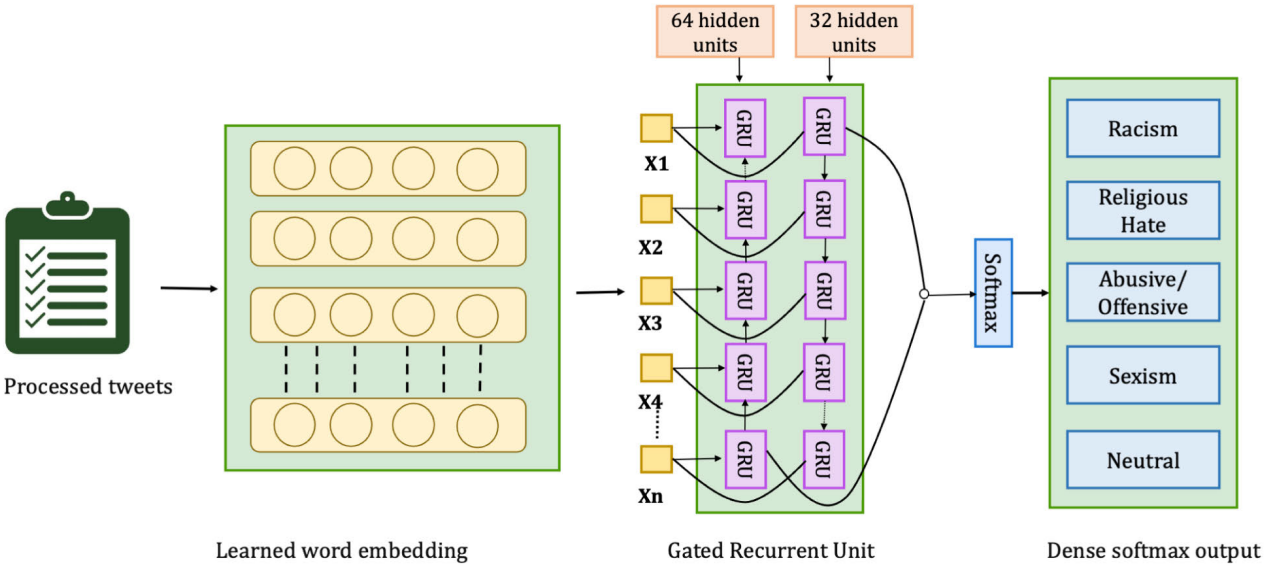

FIGURE 5. GRU model for cyberbullying detection in Roman Urdu.

TABLE 5. Results of traditional machine learning models.

ML Classifiers	Features	Accuracy	Precision	F1-Score	Recall
3*Logistic Regression	TF-IDF	89%	89.7%	88%	89.02%
	word2vec	64.2%	60.3%	57.6%	64.3%
	FastText	71.6%	70.5%	70.3%	71.6%
3*SVM	TF-IDF	94.6%	94.7%	94.5%	94.6%
	word2vec	62.3%	61.3%	53.6%	62.3%
	FastText	71%	69.8%	70%	71%
3*Random Forest	TF-IDF	95.2%	95.3%	95.1%	95.2%
	word2vec	79%	79.6%	77.8%	79%
	FastText	73.8%	75.9%	71.7%	73.8%
3*Decision Tree	TF-IDF	97.3%	97.3%	97.36%	97.37%
	word2vec	71.38%	71.5%	71.4%	71.3%
	FastText	67.5%	67.5%	67.53%	67.51%
2*Gradient Boosting	TF-IDF	95.2%	95.4%	95.2%	95.29%
	word2vec	71.1%	71%	68.75%	71.16%
2*Naive Baye's	TF-IDF	76.2%	79.2%	71.1%	76.2%
	word2vec	56.7%	32.3%	41%	56.73

until reaching a peak of 97 percent in the final epoch. Notably, this achieved accuracy surpasses that of other machine learning models utilized as binary classifiers in previous studies. For fair evaluation, the same testing part of the dataset was used for both traditional machine-learning models and deep-learning model evaluation. The performance of these trained models was tested using well-known metrics such as precision, accuracy, recall, and F1-measure, and the results are presented in Table 6. The GRU-based approach has demonstrated superior performance compared to traditional models in terms of all evaluation metrics, using word2vec embedding, as illustrated in Figure 8. Notably, among the traditional machine learning models, the Decision Tree model exhibits notably higher performance when utilizing TF-IDF as a feature extraction and word embedding technique, as shown in the comparison chart of Figure 7.

Referring to the information presented in Table 1, previous studies have highlighted that the SVM with Count Vectors and TF-IDF model outperforms both traditional and deep learning models in binary classification tasks, as assessed by accuracy, precision, recall, and f-measure. However, it is noteworthy that the performance of the SVM degrades when applied to multi-class classification of tweets. Conversely, the Bi-LSTM with attention layer (a deep learning model) performs better in the context of multi-classification of tweets, as reported in the study by Bilal et al. [11]. In our experimental findings, the GRU-based deep learning model showcased superior performance to machine learning models, including SVM, Logistic Regression, CNN, and Bi-LSTM, particularly when combined with lexical normalization of the dataset, as demonstrated in Table 6 and 7.

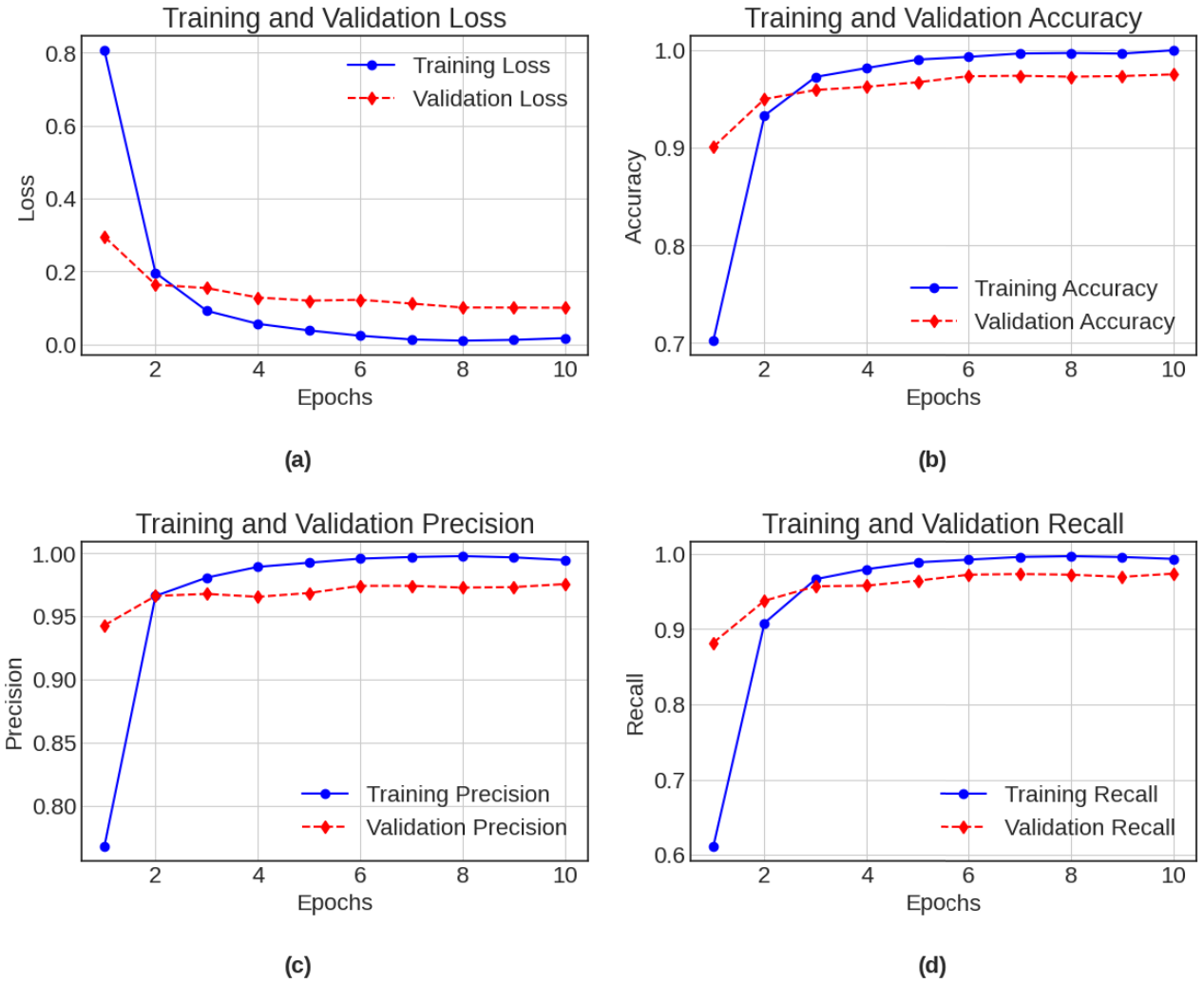

FIGURE 6. Performance of GRU for each epoch.

TABLE 6. Results of GRU and traditional machine learning models using word2vec embedding.

Classifier Model	Accuracy	Precision	F1-Score	Recall
Logistic Regression	64.2%	60.3%	57.6%	64.3%
Decision Tree	71.38%	71.5%	71.4%	71.3%
Naive Baye's	56.7%	32.3%	41%	56.73
Random Forest	79%	79.6%	77.8%	79%
SVM	62.3%	61.3%	53.6%	62.3%
Gradient Boosting	71.1%	71%	68.75%	71.16%
GRU based approach (ours)	**97.51%**	**97.55%**	**97.4%**	**97.37%**

C. ABUSER PROFILE IDENTIFICATION

The Roman Urdu tweets data, which has already been classified by the proposed model with an accuracy of 97%, is leveraged to detect user behavior. The classification of users into normal and abusive categories followed these steps:

- **Bullying Behaviour (BB):** This is calculated by determining the total number of cyberbullying tweets (Sexism and Abusive/Offensive) sent by a user, denoted as X.
- **Normal Behaviour (NB):** This is calculated by determining the total number of tweets (categories other than Sexism and Abusive/Offensive) sent by user X.
- **Total Number of Tweets (NT):** This represents the total number of tweets sent by a user.

For a user X, the probability of BB, P_{BB} and the probability of NB, P_{NB} are calculated using equations 4 and 5, respectively:

$$P_{BB} = \frac{BB(X)}{NT(BB(X) + NB(X)} \tag{4}$$

$$P_{NB} = 1 - P_{BB} \tag{5}$$

After determining the probabilities based on the user's tweets, abuser and normal profiles are identified

TABLE 7. Comparison of the proposed approach with multi-class classification model of existing studies.

Model	Features	Accuracy	Precision	F1-Score	Recall
2*Bi-LSTM [11]	word2vec	67%	68%	67%	67%
	FastText	64%	64%	64%	64%
2*CNN [16]	BERT	82%	75%	74%	75%
	FastText	66%	45%	41%	42%
GRU based approach (ours)	word2vec	95.2%	95.4%	95.2%	95.29%

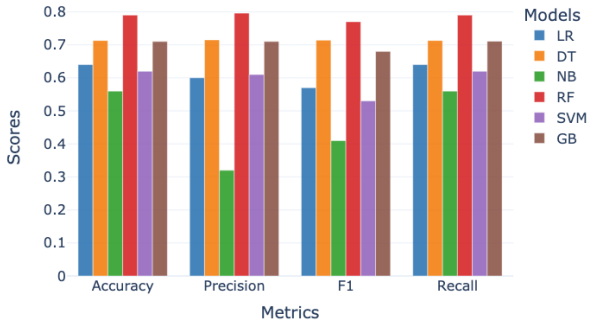

FIGURE 7. Performance of ML models using word2vec. LR = Logistic Regression, DT = Decision Tree, NB = Multinomial Naive Baye's, RF = Random Forest, SVM = Support Vector Machine, GB = Gradient Boosting.

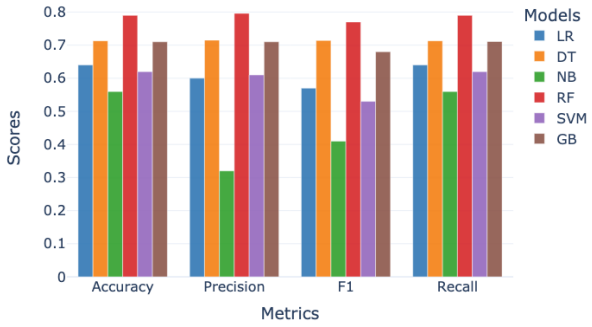

FIGURE 8. Performance of GRU and traditional ML models using word2vec embedding.

using three rules derived from the study by Sarna and Bhatia [23]:

1) If $P_{BB} < 0.5$, the user is categorized as a **Normal User**, indicating that more than half of their activities are normal.
2) If $0.5 \leq P_{BB} < 0.6$ and NT > 1, the user is classified as a **Suspicious User**, warranting a warning due to their bullying behavior being close to that of Abusive Users.
3) If $P_{BB} \geq 0.6$ and NT > 1, the user is considered an **Abusive User**, as they exhibit intensive cyberbullying behavior towards others.

Following the classification process, the model effectively categorized the data into multiple classes. Subsequently, the users' behavior and credibility were assessed using the probabilistic model described earlier. As a result, the system proficiently identified cyberbullying actors, comprising 922 Normal Users, 57 Suspected Users, and 25 Abusive Users, as depicted in Figure 9. These findings indicate the presence of cyberbullying tweets within the dataset, and further reveal that 82 users should be alerted regarding their usage of abusive or offensive language in their tweets.

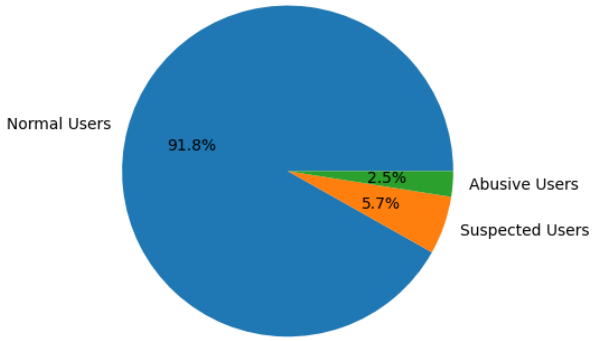

FIGURE 9. Result of abusive profile identification.

VI. CONCLUSION AND FUTURE WORK

The findings of this study underscore the critical challenges faced by Roman Urdu in the context of cyberbullying content detection, primarily stemming from the limited availability of comprehensive datasets containing user profile information. This scarcity imparts significant implications for the accurate identification and categorization of abusive content spanning diverse cyberbullying categories in the Roman Urdu text on social media. The proposed method, utilizing the GRU model with pre-learned embeddings, exhibited remarkable advancements over conventional machine learning and deep learning models, including LSTM, Bi-LSTM with attention layers, and CNN. These advancements were observed across key performance metrics such as accuracy, precision, and F-measure, with the GRU-based method attaining an accuracy rate of 97%, outperforming alternative methods in this domain. Furthermore, our investigation highlighted the effectiveness of employing Decision Tree and Gradient Boosting as conventional machine learning classifiers in conjunction with TF-IDF for cyberbullying detection. These classifiers demonstrated relatively superior performance, potentially attributed to the incorporation of lexical normalization during the data pre-processing phase. This standardization of Roman Urdu words contributed to improved performance and accuracy. The proposed method exhibited the capability to successfully classify users into three distinct categories based on their cyberbullying behavior: Normal Users (922 instances), Suspected Users (57 instances), and Abusive Users (25 instances). This categorization provides

a pivotal foundation for subsequent research and proactive measures to address cyberbullying issues in the Roman Urdu context. In future endeavors, the proposed model holds promise for implementation in various social media platforms, online safety initiatives, and content moderation systems, providing automated identification and mitigation of cyberbullying incidents. Additionally, this framework can contribute value to community management, parental control tools, research studies, and educational awareness programs, thereby fostering safer online environments and facilitating proactive measures against bullying behavior. To augment the model's capabilities, future investigations should consider the incorporation of smileys and emojis into the dataset. This addition has the potential to enhance the prediction and detection of bullying content by capturing the underlying emotional context within the tweets. Moreover, addressing inter-rater reliability and minimizing labeling discrepancies becomes imperative as data labeling may involve subjective interpretation in categorizing bullying behavior. Ensuring consistency in labeling is essential to refine the model's performance and enhance its accuracy. Furthermore, in the context of abusive profile identification, augmenting the dataset with additional features beyond tweets, such as the social network of users, daily tweet frequency, and the use of hashtags or mentions, could yield more precise and accurate results. Expanding the feature set enables a comprehensive profiling of users engaged in abusive behavior, facilitating more effective detection and mitigation strategies.

ACKNOWLEDGMENT

Open Access funding provided by the Qatar National Library. This work was supported by the Higher Education Commission (HEC), Pakistan, through the National Research Program for Universities (NRPU) under Grant 20-16597/NRPU/R&D/HEC/2021.

REFERENCES

[1] M. Woodward. (Jan. 2023). *Social Media in Pakistan—2023 Stats Platform Trends*. Accessed: Apr. 15, 2023. [Online]. Available: https://oosga.com/social-media/pak/

[2] T. Agrawal and V. D. Chakravarthy, "Cyberbullying detection and hate speech identification using machine learning techniques," in *Proc. 2nd Int. Conf. Interdiscipl. Cyber Phys. Syst. (ICPS)*, May 2022, pp. 182–187.

[3] D. Aditya, S. Kalaskar, O. Kumbhar, and R. Dhumal, "Cyber bullying detection on social media using machine learning," in *Proc. ITM Web Conf.*, vol. 40, 2021, pp. 30–38, doi: 10.1051/itmconf/20214003038.

[4] U. Khan, S. Khan, A. Rizwan, G. Atteia, M. M. Jamjoom, and N. A. Samee, "Aggression detection in social media from textual data using deep learning models," *Appl. Sci.*, vol. 12, no. 10, p. 5083, May 2022.

[5] B. A. H. Murshed, J. Abawajy, S. Mallappa, M. A. N. Saif, and H. D. E. Al-Ariki, "DEA-RNN: A hybrid deep learning approach for cyberbullying detection in Twitter social media platform," *IEEE Access*, vol. 10, pp. 25857–25871, 2022.

[6] P. Hajibabaee, M. Malekzadeh, M. Ahmadi, M. Heidari, A. Esmaeilzadeh, R. Abdolazimi, and J. H. J. Jones, "Offensive language detection on social media based on text classification," in *Proc. IEEE 12th Annu. Comput. Commun. Workshop Conf. (CCWC)*, Jan. 2022, pp. 92–98.

[7] H. Ali, A. Ullah, T. Iqbal, and S. Khattak, "Pioneer dataset and automatic recognition of Urdu handwritten characters using a deep autoencoder and convolutional neural network," *Social Netw. Appl. Sci.*, vol. 2, no. 2, pp. 1–12, Feb. 2020.

[8] H. Ali, K. Iqbal, G. Mujtaba, A. Fayyaz, M. F. Bulbul, F. W. Karam, and A. Zahir, "Urdu text in natural scene images: A new dataset and preliminary text detection," *PeerJ Comput. Sci.*, vol. 7, p. e717, Sep. 2021.

[9] F. Rasheed, M. Anwar, and I. Khan, "Detecting cyberbullying in Roman Urdu language using natural language processing techniques," *Pakistan J. Eng. Technol.*, vol. 5, no. 2, pp. 198–203, Sep. 2022.

[10] A. M. Alduailaj and A. Belghith, "Detecting Arabic cyberbullying tweets using machine learning," *Mach. Learn. Knowl. Extraction*, vol. 5, no. 1, pp. 29–42, Jan. 2023.

[11] M. Bilal, A. Khan, S. Jan, and S. Musa, "Context-aware deep learning model for detection of Roman Urdu hate speech on social media platform," *IEEE Access*, vol. 10, pp. 121133–121151, 2022.

[12] B. A. Chandio, A. S. Imran, M. Bakhtyar, S. M. Daudpota, and J. Baber, "Attention-based RU-BiLSTM sentiment analysis model for Roman Urdu," *Appl. Sci.*, vol. 12, no. 7, p. 3641, Apr. 2022.

[13] R. Ghosh, B. T. Student, S. Nowal, and G. Manju, "Social media cyberbullying detection using machine learning in Bengali language," *J. Int. J. Eng. Res. Technol.*, vol. 10, no. 5, pp. 190–193, 2021.

[14] M. M. Khan, K. Shahzad, and M. K. Malik, "Hate speech detection in Roman Urdu," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 20, no. 1, pp. 1–19, Jan. 2021.

[15] K. Sreelakshmi, B. Premjith, and K. P. Soman, "Detection of hate speech text in Hindi-English code-mixed data," *Proc. Comput. Sci.*, vol. 171, pp. 737–744, Jan. 2020.

[16] H. Rizwan, M. H. Shakeel, and A. Karim, "Hate-speech and offensive language detection in Roman Urdu," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Jan. 2020, pp. 2512–2522.

[17] H. Nurrahmi and D. Nurjanah, "Indonesian Twitter cyberbullying detection using text classification and user credibility," in *Proc. Int. Conf. Inf. Commun. Technol. (ICOIACT)*, Mar. 2018, pp. 543–548.

[18] E. A. Abozinadah, A. V. Mbaziira, and J. H. J. Jones, "Detection of abusive accounts with Arabic tweets," *Int. J. Knowl. Eng.*, vol. 1, no. 2, pp. 113–119, 2015.

[19] Z. Zhang, D. Robinson, and J. Tepper, "Detecting hate speech on Twitter using a convolution-GRU based deep neural network," in *The Semantic Web*, A. Gangemi, R. Navigli, M.-E. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai, and M. Alam, Eds., Cham, Switzerland: Springer, 2018, pp. 745–760.

[20] B. Bhatia, A. Verma, and R. Katarya, "Analysing cyberbullying using natural language processing by understanding jargon in social media," in *Proc. ICSAC*, 2021, pp. 397–406.

[21] J. J. Dooley, J. Pyzalski, and D. Cross, "Cyberbullying versus face-to-face bullying: A theoretical and conceptual review," *J. Psychol. Psychotherapy*, vol. 9, no. 3, pp. 1–11, 2019.

[22] J. W. Patchin and S. Hinduja, "Cyberbullying among adolescents: Implications for empirical research," *J. Adolescent Health*, vol. 53, no. 4, pp. 431–432, Oct. 2013.

[23] G. Sarna and M. P. S. Bhatia, "Content based approach to find the credibility of user in social networks: An application of cyberbullying," *Int. J. Mach. Learn. Cybern.*, vol. 8, no. 2, pp. 677–689, Apr. 2017.

[24] R. M. Kowalski, G. W. Giumetti, A. N. Schroeder, and M. R. Lattanner, "Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth," *Psychol. Bull.*, vol. 140, no. 4, p. 1073, 2014.

[25] 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.

AYESHA ATIF received the master's degree in computer science from the University of Engineering and Technology Lahore. In terms of professional experience, she is currently a Senior Software Engineer with i2c Inc., where she is a Web Developer in Java. Her research interests include machine learning, artificial intelligence, and NLP.

AMNA ZAFAR received the Ph.D. degree in computer science from UET Lahore, in 2019. She is currently an Assistant Professor with the Department of Computer Science, UET Lahore. She has published numerous research articles in internationally reputed journals. Her research interests include wireless IoT, e-health, artificial intelligence, and machine learning.

MUHAMMAD WASIM has a diverse industry-academia experience spanned more than 18 years. He worked on many national and international projects as an Industry Professional. From an academic perspective, he has worked with many organizations, including KICS, UET Lahore, FC College Lahore, and the University of Management and Technology (UMT). He also won three HEC-funded research projects one as the PI and two as the Co-PI related to natural language processing and machine learning. He is currently an Assistant Professor with the Computer Science Department, UMT, Sialkot Campus. His research interests include natural language processing, information retrieval, and deep learning.

TALHA WAHEED received the B.S. degree in CS and the M.S. degree in AI from NUCES Lahore, the M.A. degree in philosophy from Punjab University, Lahore, and the Ph.D. degree in knowledge management of Unani medicines from UET. He is currently an Assistant Professor with the Department of CS, UET Lahore. He mostly teaches courses, such as automata theory, compilers, AI, argument, and reasoning. He spent three decades in academia, the software industry, and the herbal industry. He is a bibliophile and belongs to a family renowned for Unani Medicine knowledge sharing and practices in the sub-continent for generations. His research interests include traditional medicines informatics, Quran informatics, e-learning, activity theory, and knowledge modeling.

AMJAD ALI received the B.S. and M.S. degrees in computer science from COMSATS Institute of Information Technology, Pakistan, in 2006 and 2008, respectively, and the Ph.D. degree from the Electronics and Radio Engineering Department, Kyung Hee University, South Korea, in 2015. From 2015 to 2022, he was an Assistant Professor with the Department of Computer Science, COMSATS University Islamabad, Lahore Campus, Pakistan. He was a Postdoctoral Fellow and a Research Professor with the Department of Information and Communication Engineering, Inha University, and the School of Electrical Engineering, Korea University, South Korea, from 2018 to 2019. Currently, he is a Postdoctoral Researcher with the College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar. His main research interests include multimedia transmission, cognitive radio networks, stochastic network optimization, machine learning, the Internet of Things, cloud and edge computing, 6G networks, and metaverse.

HAZRAT ALI (Senior Member, IEEE) is currently a Researcher in generative artificial intelligence and image processing. He is an Associate Editor of IEEE, a Book Editor of Springer, and has served as a reviewer for *Scientific Reports* (Nature), IEEE TRANSACTIONS ON ARTIFICIAL INTELLIGENCE, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, and IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE Conference on Artificial Intelligence, and many other reputed journals and conferences. He was selected as a Young Researcher of the 5th Heidelberg Laureate Forum, Heidelberg, Germany.

ZUBAIR SHAH received the M.S. degree in computer system engineering from Politecnico di Milano, Italy, and the Ph.D. degree from the University of New South Wales, Australia. He was a Research Fellow with Australian Institute of Health Innovation, Macquarie University, Australia, from 2017 to 2019. He is currently an Assistant Professor with the Division of ICT, College of Science and Engineering, HBKU. His expertise is in the field of artificial intelligence and big data analytics, and their application to health informatics. He has published his work in various A-tier international journals and conferences. His research interests include health informatics, particularly in relation to public health, using social media data (e.g., Twitter) and news sources to identify patterns indicative of population-level health.

Open Access funding provided by 'Qatar National Library' within the CRUI CARE Agreement