IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 5 June 2024, accepted 21 June 2024, date of publication 27 June 2024, date of current version 8 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3420131

RESEARCH ARTICLE

Toward Multi-Modal Approach for Identification and Detection of Cyberbullying in Social Networks

MAHMOUD AHMAD AL-KHASAWNEH[1,2,3], MUHAMMAD FAHEEM[4], (Member, IEEE), ALA ABDULSALAM ALAROOD[5], SAFA HABIBULLAH[5], AND EESA ALSOLAMI[5]

[1]School of Computing, Skyline University College, University City Sharjah, Sharjah, United Arab Emirates
[2]Applied Science Research Center, Applied Science Private University, Amman 11931, Jordan
[3]Jadara University Research Center, Jadara University, Irbid 21110, Jordan
[4]School of Technology and Innovations, University of Vaasa, 65200 Vaasa, Finland
[5]College of Computer Science and Engineering, University of Jeddah, Jeddah 21959, Saudi Arabia

Corresponding author: Muhammad Faheem (muhammad.faheem@uwasa.fi)

The work of Muhammad Faheem was supported in part by the Academy of Finland and University of Vaasa, Finland.

ABSTRACT Given the widespread use of social networks in people's everyday lives, cyberbullying has emerged as a major threat, especially affecting younger users on these platforms. This matter has generated significant societal apprehensions. Prior studies have primarily concentrated on analyzing text in relation to cyberbullying. However, the dynamic nature of cyberbullying covers many goals, communication platforms, and manifestations. Conventional text analysis approaches are not effective in dealing with the wide range of bullying data seen in social networks. In order to tackle this difficulty, our suggested multi-modal detection approach integrates data from diverse sources including photos, videos, comments, and temporal information from social networks. In addition to textual data, our approach employs hierarchical attention networks to record session features and encode various media information. The resulting multi-modal cyberbullying detection platform provides a comprehensive approach to address this emerging kind of cyberbullying. By conducting experimental analysis on two actual datasets, our framework exhibits greater performance in comparison to many state-of-the-art models. This highlights its effectiveness in dealing with the intricate nature of cyberbullying in social networks.

INDEX TERMS Cyberbullying, multi-modality, social media, hierarchy attention.

I. INTRODUCTION

Young people mostly utilize social networking as their main platform for social engagement. However, the widespread prevalence of cyberbullying on digital platforms presents a significant danger to the welfare of young individuals [1]. Over 40% of American teenagers have experienced cyberbullying on social media, according to data from the White House and the American Psychological Association [2]. Recent British research has highlighted the extent of the problem, revealing that cyberbullying is more common than bullying in real-life situations. Specifically, 12% of students reported experiencing cyberbullying. The prevalence of social cyberbullying occurrences is increasing annually, transforming into a complex dilemma that involves various objectives, routes, and manifestations. The rising problem has caused significant adverse effects on the physical and mental well-being of the victims, leading some to contemplate suicide [3]. In addition to being a distressing experience for individuals, it has evolved into a significant public health issue, leading to an increase in research efforts in the fields of psychology and computer science. This study is to understand the attributes of cyberbullying and, ultimately, to find efficient approaches to detect and tackle instances of bullying on social networks.

In the domain of automated cyberbullying identification, where there is a high occurrence of harmful verbal attacks, current efforts mostly focus on examining textual characteristics. Several methods have been created to classify text and identify instances of cyberbullying. Cyberbullying is

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

defined as when people or groups repeatedly post offensive or violent content on social media using embedded devices with the intention of hurting or upsetting other people. However, depending exclusively on the examination of textual characteristics, one faces difficulties in determining if the content is aimed at certain persons or groups without contextual information. Moreover, the existence of unsuitable visual information within conventional text-based material presents a potential hazard on social media platforms. Hence, it is crucial to highlight the significance of essential information contained in diverse types of social media, such as photographs, videos, comments, and social networks.

Present efforts centered around multi-modal information tend to give priority to particular modalities. Comments are commonly perceived as concise exchanges over a specific topic. The study [4] utilized contextual information to improve the understanding of the whole context and the determination of conduct. Nevertheless, this study failed to consider the interplay between individual comments while striving to comprehend the correlation between comments. Soni [5] proposed an alternate method that incorporated visual attributes to overcome the constraints of textual elements. Although these methods demonstrate improved performance in comparison to text analysis alone, they are not effective in overcoming the limitations associated with single-mode information. Additionally, important traits like persistence and the gradual recurrence of hostile acts are displayed by cyberbullying [2]. Interrupting conversations about cyberbullying while successfully averting secondary damage presents a new problem. Therefore, a new difficulty in cyberbullying detection is identifying multi-modal bullying material quickly enough to stop further discussion.

We reinterpret the phenomena as a method that uses textual, visual, and extra meta information to determine whether a post is related to a bullying topic in response to these changing forms of cyberbullying. Our novel Multi-Modal Cyberbullying Detection (MMCD) framework is introduced to address the aforementioned issues. To consistently recognize various instances of cyberbullying on social networks, this framework combines textual, visual, and other meta-information. In particular, we propose that offensive remarks be made on blogs that engage in cyberbullying. Utilizing Hierarchical Attention Networks (HAN) [6], we evaluate the significance of each comment by modeling it and then encoding visual and other meta-data. To improve cyberbullying detection performance, these traits and textual content are combined. Among this work's principal contributions are:

- We provide an innovative view of the complete problem-solving process that is based on the combination and combined processing of multi-modal data to successfully handle the various types of cyberbullying.
- We developed an original multi-modal framework for cyberbullying detection. This framework employs an innovative approach by independently modeling textual, visual, and other information. It incorporates a self-attention-based BiLSTM model, a HAN model focusing on word and comment levels, and other embedding techniques. The integration of these components enables efficient information merging, contributing to the effective resolution of the complex issue of cyberbullying.
- We acquired data based on multi-modality from prominent social media website, namely Twitter and Facebook, to validate the efficacy of our approach. Additionally, we conduct a thorough investigation into the impact of multi-modal data on cyberbullying.

II. LITERATURE REVIEW

A significant amount of earlier work on cyberbullying detection has focused on text feature analysis as a way to identify bullying behaviors. Emotional analysis and text classification are common methods that employ N-gram models, BoW, and TF-IDF [7], [8], and [9]. Classification methods such as Random Forest, Support Vector Machine (SVM), Logistic Regression, and Naive Bayes have proven to be effective in handling these features [8], [10], [11], [12]. Specifically, Chavan and Shylaja [13] detected bullying behavior by extracting variables such as TF-IDF, BoW, and bullying lexicon using SVM and logistic regression. Aside from text features, network attributes have also been studied [14]. Academics have studied a variety of social network metrics, including the volume of tweets, geographic distribution, and the strength of users' social networks. Using the nature of social networks, Chelmis et al. [15] developed a system to detect cases of bullying. On the other hand, Algaradi et al. developed a detection model that effectively combines metrics related to networks, user behavior, and tweet content. To improve the overall performance of their system, Cheng et al. [11] designed a complex heterogeneous network that included metadata like user profiles, photos, videos, time, location, and comments. With the post-vector representations learned by network embedding, they proceeded to employ SVM and Random Forest classification techniques. With these models, we have a thorough strategy for identifying cyberbullying, and we have solved the problem of weak text features.

As an end-to-end approach, deep learning also shows improved text representation capabilities [16]. Text classification has come a long way thanks to convolutional neural networks (CNN) [17], recurrent neural networks (RNN) [18], and hybrid architectures like RCNN, which combines CNN and RNN, [19]. Miswriting is a frequent occurrence on social media and is frequently used in bullying texts to avoid discovery. Park and colleagues [20] tackled this problem by implementing a hybrid model that connected character and word convolutional neural networks in a smooth manner for efficient categorization. 21. Zhang et al. [21] came up with a new way to encode text that combined convolution layers with a Gate Recurrent Unit (GRU) to include both structural and sequence information.

Cyberbullying detection systems now use attention techniques to highlight key terms. An attention mechanism-based bidirectional RNN (BiRNN) model [22] with an was

presented by Zhang et al. [23] to detect bullying text. Word weights were adjusted using an attention mechanism in this model, which also integrated contextual data via BiRNN. Deep models, which are similar to these methods, have relied heavily on specific meta-information. By merging latent representations from text and information, Founta et al. [24] developed a hybrid model. Yafooz et al. [25] employed transfer leaning approaches to detect and classify the kids cyberbullying on social media. They used two Arabic dataset collected from YouTube Videos, then applied several pre-trained models. The best accuracy has been recorded using AraBERT model which reached to 95% and 96% using both datasets. Similarly, Alhejaili et al. [26] detecting hate speech using machine learning classifiers.

Owing to social media's diversity, some researchers have integrated visual data into their textual analysis in addition to broadening it. Soni et al. [5] made an effort to derive visual cues to fill in the gaps left by the absence of text. Li et al. [4] examined child semantics and child comments on related issues to gain a richer comprehension of context by utilizing the parent-child link between comments. Also, the study [27], which built a time-dependent hierarchical attention network to gather comment features, shows that techniques for classifying documents have been used on comments. These techniques highlight the advantages of multi-modal data for cyberbullying detection. In the topic of cyberbullying detection, current research developments include feature fusion and feature extraction from multi-modal datasets.

III. PROBLEM FORMULATION

A corpus of N social media posts is represented by $\boldsymbol{p} = \{\boldsymbol{p_1}, \boldsymbol{p_2}, \boldsymbol{p_3}, \ldots\ldots \boldsymbol{p_n}\}$. A media object $\boldsymbol{M_i}$, such as an image or video, a timestamp, user profile, Likes, and Shares, text content $(\boldsymbol{T_i})$, and additional information $(\boldsymbol{O_i})$ such as a timestamp, user profile, Likes, and Shares, comprise each post (P_i). The mathematical representation of a post $(\boldsymbol{P_i})$ is given by $\boldsymbol{P_i} = \{\boldsymbol{T_i}, \boldsymbol{C_i}, \boldsymbol{M_i}, \boldsymbol{O_i}\}$, where $\boldsymbol{T_i}$ signifies text content, $\boldsymbol{C_i}$ represents a collection of comments, $\boldsymbol{M_i}$ denotes the media object, and $\boldsymbol{O_i}$ encompasses additional meta-data.

Moreover, the length of $\boldsymbol{C_i^{(1)}}$ is indicated by $\boldsymbol{\ell_i^{(1)}}$, and $\boldsymbol{C_i^{(1)}}$ represents the first comment in $\boldsymbol{C_i}$. Furthermore, every post $\boldsymbol{P_i}$ is given a binary label $\boldsymbol{Y_i} = \{\boldsymbol{0}, \boldsymbol{1}\}$, where 1 signifies bullying conduct and 0 suggests the opposite. In our dataset representation, the length of comments in each post $\boldsymbol{P_i}$ is determined by the total number of comments in the collection $\boldsymbol{C_i}$. The length of the initial comment $\boldsymbol{C_i}$ is indicated as $\boldsymbol{\ell_i^{(1)}}$, with $\boldsymbol{C_i^{(1)}}$ representing the first comment in the collection $\boldsymbol{C_i}$. This notation enables us to distinguish and examine individual comments within each post, offering valuable insight into the structure and content of the comments. In our cyberbullying detection framework, the symbol $\boldsymbol{\ell_i^i}$ acts as a reference to accurately measure and analyze the characteristics of the first comment's length in the collection.

We present a function $\boldsymbol{F}$ that learns bullying behavior by taking into account the context, comments, media, and other pertinent data in order to formalize the cyberbullying detection process. This can be stated as follows:

$$\boldsymbol{P_i} = \{\boldsymbol{T_i}, \boldsymbol{C_i}, \boldsymbol{M_i}, \boldsymbol{O_i}\}$$

In this context, the binary label denoted as $\boldsymbol{Y_i}$ signifies whether harassing behavior was present (1) or absent (0) in the given post Pi. The objective of the function F is to identify the correlations that exist between the different components of the post and the likelihood of cyberbullying.

The function $\boldsymbol{F}$ is subjected to training, validation, and evaluation to acquire knowledge and identify instances of bullying behavior. During the training process, labeled data instances are utilized to optimize the model parameters using techniques such as gradient descent. Afterward, the function is verified on distinct data to ensure its applicability to various scenarios and optimize the hyperparameters. Ultimately, the system's performance is assessed by measuring metrics such as accuracy and precision on a separate test dataset. This process guarantees that $\boldsymbol{F}$ accurately detects harassing behavior in various posts and situations, offering valuable information about its sensitivity and performance.

IV. MULTI-MODAL APPROACH FOR CYBERBULLYING DETECTION

In this section, we present the MMCD architecture in detail. Two separate processes, encoding and decoding, allow the model to operate. During the encoding phase, a lot of different components are used to encode different types of data. These components include a BiLSTM-based Topic-oriented encoder, a hierarchical attention mechanism based on comments, media embedding, and extra meta-information embedding layers. In addition, the comments encoder takes into account the sequential structure of the collection of comments.

The decoding procedure utilizes a multilayer perceptron (MLP) to train the multi-mode data separately. Eventually, the multi-mode data is integrated for comprehensive training. Figure 1 depicts the schematic representation of the proposed Multi-Modal Cyberbullying Detection framework.

The MMCD architecture incorporates multiple modalities, such as text, comments, media, and metadata, which are individually processed and combined to improve cyberbullying detection. The textual content is encoded using BiLSTM networks, effectively capturing both forward and backward sequential dependencies in sentences. Attention mechanisms are utilized to assess the importance of individual words, highlighting relevant information during the encoding process. Comments are processed through a hierarchical attention mechanism based on document classification. Bidirectional GRU models sequentially encode words, utilizing attention mechanisms to emphasize significant words in comments. The media data is encoded using one-hot encoding and then passed through a multi-layer perceptron to extract features efficiently. Simultaneously, metadata encompassing timestamps and user profiles are incorporated to furnish a supplementary context. Attention mechanisms play a vital role in various modalities, allowing the model to

concentrate on relevant information during encoding. In the decoding phase, the encoded representations from various modalities are combined, and fully connected units adjust the weights of vectors from each modality. The MMCD architecture utilizes a comprehensive approach to effectively analyze and integrate various modalities, thereby improving its ability to detect cyberbullying.

A. MULTI-MODAL ENCODER

BiLSTM networks hold a pivotal role in the landscape of natural language processing, specifically when confronted with the task of analyzing sequences of sentences. Long Short- Term Memory (LSTM), as a refinement of Recurrent Neural Networks (RNN), introduces three essential gated components: the input gate i_t, the forget gate f_t, and the output gate o_t. These gates' states intricately depend on the previous state h_{t-1}, where ht denotes the state at time t. Equation 1 expresses the computation of it as the result of applying the sigmoid function σ to the sum of the product of vector x_t and weight matrix W_{xi}, the product of vector h_{t-1} and weight matrix W_{hi}, and the bias vector b_i.

$$it = \sigma(x_t W_{xi} + h_{t-1} W_{hi} + b_i) \tag{1}$$

Equation 2 calculates the value of f_t by applying the sigmoid function to the total of x_t multiplied by W_{xf}, h_{t-1} multiplied by W_{hf}, and b_f.

$$f_t = \sigma(x_t Wxf + h_{t-1} W_{hf} + b_f) \tag{2}$$

Equation 3 denotes the computation of o_t by applying the sigmoid function to the sum of x_t multiplied by W_{xo}, h_{t-1} multiplied by Who, and b_o.

$$o_t = \sigma(\text{xtWxo} + h_{t-1} W_{ho} + b_o) \tag{3}$$

Consider x_t as the word embedding vector at time t. The weight matrices W_{xi}, W_{xf}, W_{xo}, W_{hi}, W_{hf}, Who correspond to the input x_t, and biases b_i, b_f, b_o are also included. The hidden layer h_t is determined by combining the candidate memory c˜t and the current value of c˜t.

To obtained the values of $\tilde{c}_t$, apply the hyperbolic tangent function to the sum of the products of x_t and W_{xc}, h_{t-1} and W_{hc}, along with b_c.

$$\tilde{c_t} = \tanh(x_t W_{xc} + h_{t-1} W_{hc} + b_c) \tag{4}$$

In Equation 4, we express the calculation of the present cell state c_t within a recurrent neural network. This involves element-wise multiplication of the forget gate f_t and the preceding cell state c_{t-1}, followed by addition to the element-wise product of the input gate it and the candidate cell state c˜t. c˜t.

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \tag{5}$$

Equation 5 expresses the computation of the hidden state h_t as the element-wise product of the output gate o_t and the hyperbolic tangent of the cell state ct.

$$h_t = o_t * \tanh(c_t) \tag{6}$$

where the matrices in Equation 6 W_{xc} and W_{hc} represent weights, while b_c denotes a bias term.

We employ BiLSTM to encode the textual content, capturing sentence attributes from both the anterior and posterior directions. The concealed state of the i-th word, represented as $\vec{h}_i$, is computed by the utilization of an LSTM model applied to the embedding vector x_i. The calculation is executed for each word i in the sentence, with i ranging from 1 to n. Equation 7 symbolizes this procedure.

$$\vec{h}_i = \text{LSTM}(x_i), \text{ for } i \text{ ranging from 1 to } n \tag{7}$$

Equation 8 represents the calculation of the hidden state in the reverse direction h_i for I ranging from n to 1.

$$\overleftarrow{h}_i = \text{LSTM}(x_i), \textit{ for } i \text{ ranging from } n \text{ to } 1 \tag{8}$$

Furthermore, a self-attention mechanism at the word level is employed to enhance the detection of negative phrases. The detailed of bidirectional LSTM for sentence modeling is presented in Algorithm 1.

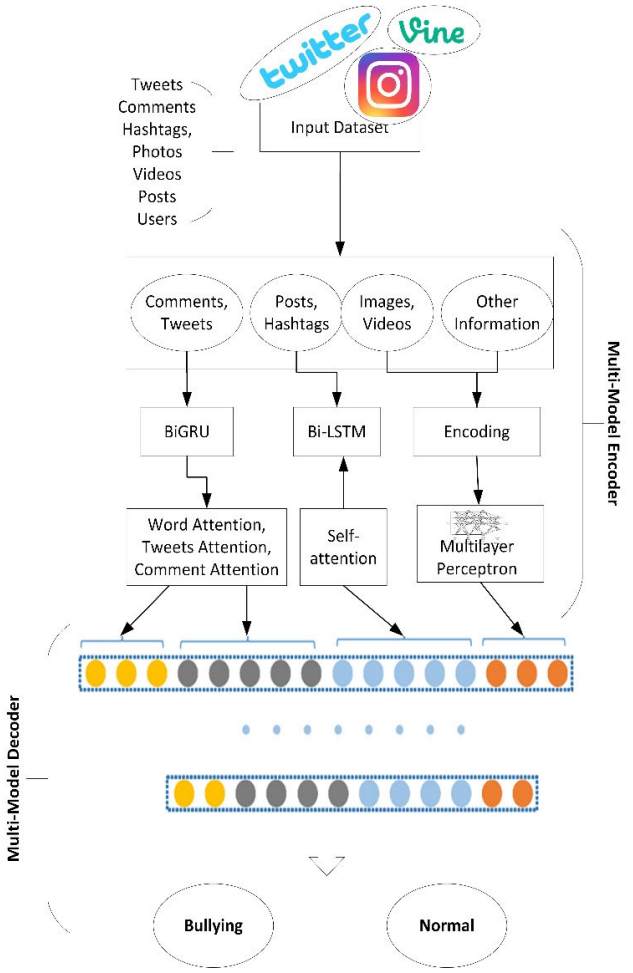

FIGURE 1. The proposed framework for multi-modal cyberbullying detection involves a rigorous analysis of social media posts. We employ various modality modeling techniques, extracting latent vectors for each modality data using a multi-modal encoder. Subsequently, these latent vectors are integrated and utilized for training classifiers through a multi-modal decoder."

B. EMBEDDING OF COMMENTS USING HIERARCHIAL ATTENTION NETWORKS

The method for analyzing and incorporating comments into the cyberbullying detection model entails several

Algorithm 1 Bidirectional LSTM for Sentence Modeling

1: **Input:** Sentence with n words: $\{w_1, w_2, \ldots, w_n\}$, Word embeddings: $\{x_1, x_2, \ldots, x_n\}$
2: **Output:** Hidden states from both directions: $\{\vec{\boldsymbol{h}}_1, \vec{\boldsymbol{h}}_2, \ldots, \vec{\boldsymbol{h}}_n\}$, $\{\overleftarrow{\boldsymbol{h}}_1, \overleftarrow{\boldsymbol{h}}_2, \ldots, \overleftarrow{\boldsymbol{h}}_n\}$
3: **for**i in $[1, n]$ **do**
4: Compute forward hidden state: $\vec{\boldsymbol{h}}_i = \vec{LSTM}(x_i)$ Equation (7)
5:Compute backward hidden state: $\overleftarrow{\boldsymbol{h}}_i = \overleftarrow{LSTM}(x_i)$ Equation (8)
6: **end for**
7: **function** LSTM(x_t, h_{t-1})
8: Compute input, forget, and output gates:
$i_t = \sigma(x_t W_{xi} + h_t - 1W_{hi} + b_i)$,
$f_t = \sigma(x_t W_{xf} + h_t - 1W_{hf} + b_f)$,
$o_t = \sigma(xtWxo + h_t - 1W_{ho} + bo)$
Equations (1), (2), (3) 9: Compute candidate memory and current memory
9: Compute candidate memory and current memory:
$\tilde{c}_t = \tanh(x_t W_{xc} + h_t - 1W_{hc} + b_c)$, $c_t = f_t * c_t - 1 + i_t * \tilde{c}_t$ Equations (4), (5) 10: Compute hidden state:
$h_t = o_t * \tanh(c_t)$ Equation (6)
11: **return** h_t
12: **end function**

essential steps. Initially, comments undergo preprocessing, including tokenization and embedding techniques, to standardize and enhance their representation. Each word within a comment is embedded using an embedding matrix, and bidirectional GRU models are employed to encode the words sequentially, resulting in hidden vectors representing the comments at the word level. Subsequently, a hierarchical attention mechanism is applied to capture the contextual significance of words within comments. This mechanism evaluates the importance of each word in the context of the entire comment, enabling the model to focus on salient words while encoding comment-level information. Additionally, an MLP with a hidden layer further refines the representation of the hidden vectors obtained from the attention mechanism, enhancing the model's ability to capture intricate relationships between words and generate a comprehensive representation of the comments. The hierarchical attention encoding process is formalized through an algorithmic approach, delineating the sequential steps involved in embedding and encoding comments using the hierarchical attention mechanism. Through these steps, our method systematically analyzes and integrates comments into the cyberbullying detection model, facilitating effective capture of the nuanced structure and content of social media comments.

It is presumed that every comment is impacted by all preceding comments. Hence, we consider all comments as a document produced in the chronological sequence of their publication. In order to encode the comments, we utilize the Hierarchical Attention Network(HAN) [6], which is a method based on document classification. This work highlights the significance of the attention mechanism at both the word level and comments level. This enables the HAN model to prioritize significant content throughout the encoding of the document. We employ word-level attention for each comment and subsequently utilize attention mechanisms at the comment level, processing the comments with a hierarchical attention architecture. We employ a Bidirectional Gated Recurrent Unit (GRU) to encode information at both the word-level and comment level. GRU, like LSTM, is a form of RNN that consists of two gates: the update gate and the reset gate. Let C be a set of comments, where L is the total number of comments. Let c_i be the i-th comment in C, consisting of L_i words denoted as w_{it}, where t ranges from 1 to L_i. To represent these words, we use an embedding matrix and compute the embedding x_{ij} as the product of the word embedding We_{wij}. Next, we feed the words of comment i into the Bidirectional GRU model for the purpose of encoding:

$$\vec{\boldsymbol{h}}_{it} = \text{GRU}(x_{it}), \quad t \in [1, L_i], \quad i \in [1, L] \tag{9}$$

$$\overleftarrow{\boldsymbol{h}}_{it} = \text{GRU}(x_{it}), \quad t \in [L_i, 1], \quad i \in [1, L] \tag{10}$$

where the hidden state from word w_{i1} to w_{iLi} is denoted by $\vec{\boldsymbol{h}}_{it}$, whereas the backward hidden state is represented by $\overleftarrow{\boldsymbol{h}}_{it}$. We combine these hidden vectors in both forward and backward directions by concatenating them: $h_i = [\vec{\boldsymbol{h}}_{it}, \overleftarrow{\boldsymbol{h}}]$.

Given that each word exerts a distinct impact on the comment, the proposed model aims to reconstruct the vector of these words using the attention mechanism in order to generate comments specifically for the significant words. More precisely, we employ a multi-layer perceptron that includes a hidden layer to extract a more advanced hiddenlayer representation.

$$u_{it} : u_{it} = \tanh(W_w h_{it} + b_w) \tag{11}$$

where W_w represents the weight matrix and b_w denotes the bias at the word-level. The similarity between u_{it} and a vector u_w, which represents a word-level context, is quantified Next, we standardize the weight matrix:

$$u_{it} = \frac{\exp(u_{it}.u_w)}{\sum_{i=1}^{L_i} \exp(u_{ij}.u_w)} \tag{12}$$

The details of hierarchical attention encoding are presented in 2.

C. ALTERNATIVE EMBEDDING APPROACH

Our first step is to generate a one-hot encoding using media tags such as text, scenery, portraits, and others to encode media-related information. To reduce the complexity of the one-hot encoding, we utilize a multi-layer perceptron to extract features efficiently.

D. MULTI-MODAL DECODER

We propose that various modalities have distinct contributions to the identification of bullying conduct. Expanding on this idea, we present a new method to modify the weights of the vectors of different modalities throughout the decoding stage.

Algorithm 2 Hierarchical Attention Encoding for Comments

1: **Input:** Comments C with L comments, each comment c_i having L_i words w_{it}, $t \in [1, L_i]$ 2: **Output:** Hierarchically encoded comments
3: **for** $i \leftarrow 1$ to L**do**
4: **for** $t \leftarrow 1$ to L_i **do**
5: Embed each word: $x_{it} = We_{wij}$
6: Encode words bidirectionally:
7: $\vec{h}_{it} = \text{GRU}(x_{it}), t \in [1, L_i]$
8: $\overleftarrow{h}_{it} = \text{GRU}(x_{it}), t \in [L_i, 1]$
9: Concatenate hidden vectors: $h_i = [\vec{h}_{it}, \overleftarrow{h}_{it}]$
10: **end for**
11: **for**$t \leftarrow 1$ to L_i **do**
12: Extract higher-level representation: $u_{it} = \tanh(W_w h_{it} + b_w)$
13: Measure similarity: $\alpha_{it} = \frac{\exp(i_{it}.u_w)}{\sum_{j=1}^{L_i} \exp(i_{ij}.u_w)}$
14: **end for**
15: **end for**

In the encoding phase, we retrieve data from several modalities and represent them using vectors of varying dimensions: v_t for comments, v_c for text, v_m for media, and v_o for other meta-information. During the decoding process, we utilize separate fully connected units for each modality and compute the values for each layer:

$$h_{dt} = tanh(W_{dt} v_t + b_{dt}), \tag{13}$$

$$h_{dc} = tanh(W_{dc} v_c + b_{dc}), \tag{14}$$

$$h_{dm} = tanh(W_{dm} vm + b_{dm}), \tag{15}$$

$$h_{do} = tanh(W_{do} v_o + b_{do}), \tag{16}$$

where the hidden layers of the completely linked units are denoted as h_{dt}, h_{dc}, h_{dm}, and h_{do}. The variables W_{dt}, W_{dc}, W_{dm}, W_{do} correspond to the weight matrices, whereas b_{dt}, b_{dc}, b_{dm}, b_{do} indicate the biases. Consequently, we resize each concealed vector to match the significance of the modalities' information. The computation of each hidden vector's output is as follows:

$$\text{out}^{\sim} = \tanh(\text{Wh} + \text{b}), \tag{17}$$

Here, W represents the weight matrix and b represents the bias for each hidden vector. The output vectors obtained are out_t, out_c, out_m, and out_o. The act of combining these separate output vectors yields the result $out = [out_t; out_c; out_m; out_o]$. The ultimate decoding procedure is subsequently executed using the vector out. The multi-model encoder formulation is presented in Algorithm 3.

V. IMPLEMENTATION ENVIRIONMENT

The study aims to evaluate and categorize content sourced from Instagram, Vine, and Twitter. The hardware infrastructure comprises an Intel Core i7-6700 central processing unit (CPU) with 18 gigabytes of random access memory (RAM), working on the Windows 10 operating system. The major programming language used is Python version 3.6, and development takes place in the Visual Studio

Algorithm 3 Multi-Modal Decoder

1: Input: Encoded vectors for comments, text, media, and other meta-information:
vt, vc, vm, vo
2: Output: Decoded vector: out
3: for each modality do
4: Compute hidden layer with fully connected units:
$h_{dmod} = \tanh(W_{dmod}$ v mod $+ b_{dmod})$
5: end for
6: Adjust the size of each hidden vector:
$\text{out}^{\sim} = \tanh(\text{Wh} + \text{b})$
7: Concatenate output vectors:
$\text{out} = [\text{out}_t; \text{out}_c; \text{out}_m; \text{out}_o]$
8: Return: Final decoded vector out

Code (VS Code) integrated development environment. The project's technological stack includes essential libraries such as Matplotlib for data visualization, NLTK for text analysis, Keras for neural network development, Pandas for efficient data preparation, Sklearn for machine learning classification, and TweetInvi for interfacing with the Twitter API. This comprehensive framework guarantees efficient analysis and categorization of various information across these social media networks. Table 1, summarizes the implementation environment of the proposed model.

TABLE 1. Implementation environment for the proposed multi-model approach for cyberbullying detection.

Component	Tools and Technologies	Description
Used Hardware	Programming language	Python version 3.6
	IDE	VS Code 1.74 V
	RAM	18 GB
	OS	Windows 10
	CPU	Intel (R) Core (TM) i7-6700 CPU 3.40 GHz
Core Libraries	Matplotlib	Data visualization libraries
	NLKT	Text analysis support toolkit and libraries
	Keras	Neural network libraries support
	Pandas	Data preparation pandas libraries
	Sklearn	Machine learning support library for classification
	TweetInvi	Twitter API support library
	Instagram	Instagram API support library for data retrieval and analysis
	Vine	Vine API support library for accessing and analyzing Vine content

A. DATASET DESCRIPTION

For our experimental assessments, we utilized two datasets obtained from well-known social media platforms: Instagram, a platform that focuses on uploading photos and videos, and Twitter, a widely used service for microblogging and

social networking. The datasets provided are openly available and cover a wide range of data types, such as tweets, text, pictures, and others.

1) VINE DATASET

The dataset is referred to as the Vine dataset [28]. Vine is a social media platform that enables users to share brief, six-second, repetitive video segments. The dataset has 970 268 posts, of which 666 demonstrate normal conduct and demonstrate bullying behavior. Every Vine post comes with content, user comments, and video tags, which together provide a comprehensive dataset for research.

2) INSTAGRAM DATASET

Instagram is a widely recognized social media platform that allows users to share photographs and videos. The dataset comprises 2,218 posts, of which 678 were classified as instances of abuse and 1,540 were classified as normal [29]. Additionally, a combined total of 155,260 comments are associated with these posts. Detailed data, such as user profiles, image annotations, timestamps, and user feedback, is accessible for review

3) TWITTER DATASET

The information gathered from the popular microblogging site Twitter includes, among other things, a heterogeneous collection of tweets, hashtags, users, and locations (see references [30], [31], [32]). This dataset includes examples of both friendly and hostile behavior. Every tweet in the dataset has links to a variety of components, such as user profiles, textual content, hashtags, timestamps, and user feedback. Out of the 30,000 posts included in the dataset, 14,250 were found to contain abusive content, and the remaining 15,750 were categorized as typical.

VI. EXPERIMENTAL ANALYSIS AND RESULTS

This section evaluates our methods using experiments on three real-time datasets from Twitter, Vine, and Instagram. An 80/20 split of the data is made into training and evaluation sets. We use measurements like accuracy (ACC) and F1 scores to assess performance. Using a pre-trained Glove model to train words and encode them as 300-dimensional vectors, the process includes word embedding. Figures that are composed of only black lines and shapes. These figures should have no shades or half-tones of gray, only black and white.

A. BASELINE APPROACH

We benchmarked our cyberbullying detection technique against other baseline methodologies in order to thoroughly assess its efficacy. For this, classic machine learning algorithms like Naive Bayesian models, Random Forest, Support Vector Machines (SVMs), and Logistic Regression were used. We adjusted these models to use TF-IDF vectors for text data representation at the word and character levels in order to accommodate the variety of text forms. Furthermore, we added psychological insights from the Linguistic Inquiry Word Count (LIWC) instrument to our models.

B. RESULTS

We evaluate the effectiveness of different models on the Twitter, Instagram, and Vine datasets using accuracy and F1 scores as metrics. Given the uneven data distribution in these datasets, we prioritize evaluating F1 scores. The provided results, shown in Tables 2, highlight how well MMCD performs compared to the other models, as evidenced by its better F1 and ACC scores.

MMCD performs better in the Twitter dataset than the top baseline model, MMCD, with 2.9% higher F1 and 2.5% higher ACC scores. A significant improvement in F1 scores more than offsets the slight increase in ACC scores, which is particularly helpful in identifying instances of cyberbullying. Notably, on the Vine dataset, the MMCD model demonstrates superior performance, with a 1.9% higher F1 score and a 0.3% higher ACC score compared to its own baseline. These results underscore the advantages of our approach over previous models, emphasizing enhanced accuracy and stability in the detection of cyberbullying.

The MMCD model integrates temporal components, annotations, and the framework of social media interactions. However, it fails to consider media data, such as information associated with images and videos. On the other hand, the approach of Cheng et al. depends on manually designed textual characteristics and takes into account the occurrence of keywords in comments. However, these characteristics do have specific constraints. The results validate the efficacy of utilizing media data to detect occurrences of cyberbullying, highlighting the higher proficiency of deep learning compared to traditional approaches in extracting unique characteristics. The Hierarchical Attention Network (HAN) surpasses existing deep learning models, exhibiting exceptional performance in terms of both F1 scores and ACC scores across a wide range of datasets. This highlights the effectiveness of attention mechanisms in hierarchically encoding textual information. The findings emphasize the significance of comments and the communal element in recognizing instances of cyberbullying, with attention processes playing a pivotal role in improving the accuracy of detection. The LSTM model with attention demonstrates exceptional efficacy when applied to the Twitter dataset, outperforming the normal LSTM model. The statistically significant improvements in F1 scores by 2.7% and ACC values by 2.3% indicate notable enhancements in the model's performance. Incorporating attention processes enhances the stability and accuracy of the model. In addition, the MMCD model surpasses the HAN model by highlighting the diverse significance of posts and comments in detecting cyberbullying on social media, hence boosting the efficacy of the detection method.

Furthermore, our MMCD model exhibits greater performance in comparison to the Text-CNN model. The Text-CNN model encounters difficulties when it encounters a dearth

of sequential information in the text. Traditional approaches of identifying cyberbullying, regardless of its language attributes, face challenges in achieving a high degree of efficacy. Conversely, the Random Forest model demonstrates higher performance in comparison to other classifiers, indicating that each feature serves a unique purpose across different classifiers. This underscores the possibility of attaining superior results through the amalgamation of varied attributes

The findings of this study lend credence to the notion that the MMCD model's demonstration of the importance of attention processes and the incorporation of comprehensive features plays a significant role in increasing the efficiency of cyberbullying detection. The significance of taking into account both social and textual factors together highlights the intricacy of this undertaking within a practical social media setting, as shown in Figure 2, 3, and 4.

C. PARAMETER ANALYSIS

We explore a variety of embedding techniques throughout the training phase in order to assess the impact of various word embedding strategies on our model. First, we choose a range of pre-trained models with varying dimensions, including en-word2vec-3002, en-glove-6b-300d, en-glove-42b-300d, and en-glove-840b-300d.

The results in Figures 5 and 6 show how the size of the corpus affects how well-pre- trained word embeddings work. As the corpus size grows, the Glove model consistently gets higher F1 scores and accuracy scores.

The en-glove-840b-300d model achieves superior performance compared to the en-glove-42b-300d model on the Instagram dataset, with an improvement of 0.16% in F1 scores and 2.71% in ACC scores. The Twitter dataset also shows similar patterns, with the en-glove-840b-300d model outperforming the en-glove-42b-300d model. The en-glove-840b-300d model achieves a 0.18% increase in F1 scores and a 2.46% increase in ACC scores.

When analyzing the Vine dataset, improvements can be shown, as the en-glove-840b- 362 300d model outperforms the en-glove-42b-300d model by 0.66% in F1 scores and 1.5% in ACC scores. Once again, this pattern is observed in the Twitter dataset, where the en-glove-840b-300d model has superior performance compared to the en-glove-42b-300d model. Specifically, it exhibits improvements of 0.72% in F1 scores and 1.25% in ACC 366 values.

The larger corpus size is responsible for these benefits, as it includes a wider range of languages. The 840b model, which includes more than 10% additional pre-trained words in both datasets compared to the 42b model, is essential for improving performance. Compared to the en-word2vec-300 model, the Glove models consistently perform better on all datasets.

In comparison to the en-word2vec-300 model, the Glove models consistently demonstrate greater performance on all datasets. The consistent enhancements in the performance of the Glove model, namely the 388 en-glove-840b-300d version, emphasize its appropriateness for tasks that demand a subtle comprehension of language, such as identifying instances of cyberbullying on social media platforms like Instagram, Vine, and Twitter. The results emphasize the significance of choosing a word embedding model that uses a bigger and more diversified collection of texts. This will improve the overall effectiveness of the framework when applied to various social media situations.

TABLE 2. An analysis of F1-scores and accuracy metrics across baseline models on the Twitter, Instagram, and Vine dataset.

Sr. No	ML-Models	Methods	Twitter Dataset		Vine Dataset		Instagram Dataset	
			Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score
1	Support Vector Machine (SVM)	Char TF-IDF	0.631	0.643	0.529	0.622	0.576	0.583
2		Word TF-IDF	0.586	0.628	0.571	0.587	0.556	0.562
3		LIWC	0.653	0.665	0.638	0.672	0.623	0.597
4	Naive Bayes (NB)	Char TF-IDF	0.688	0.740	0.631	0.658	0.625	0.676
5		Word TF-IDF	0.717	0.736	0.662	0.697	0.653	0.668
6		LIWC	0.628	0.552	0.638	0.559	0.592	0.504
7	Logistic Regression (LR)	Char TF-IDF	0.635	0.647	0.612	0.596	0.594	0.583
8		Word TF-IDF	0.656	0.632	0.641	0.595	0.605	0.573
9		LIWC	0.781	0.705	0.726	0.684	0.73	0.653
10	Random Forest (RF)	Char TF-IDF	0.67	0.74	0.625	0.658	0.619	0.669
11		Word TF-IDF	0.735	0.707	0.746	0.781	0.695	0.637
12		LIWC	0.798	0.717	0.761	0.729	0.758	0.604
13	LSTM		0.827	0.743	0.783	0.641	0.791	0.613
14	LSTM with Attention		0.848	0.778	0.813	0.692	0.813	0.692
15	Text-CNN		0.822	0.708	0.761	0.674	0.781	0.643
16	HAN		0.843	0.745	0.817	0.797	0.804	0.708
17	Xu et al.		0.587	0.541	0.684	0.697	0.513	0.502
18	Lu et al.		0.8909	0.815	0.817	0.797	0.851	0.783
19	**Proposed (MMCD)**		**0.921**	**0.846**	**0.838**	**0.841**	**0.864**	**0.86**

To examine the learning rate's sensitivity and influence of the learning rate, we systematically manipulate its values and assess their impact on overall performance, with a specific emphasis on F1 scores.

Figure 7 indicates the durability of our model across a wide range of learning rates. However, it shows that the model's performance is not optimal when the learning rate is high or low. A learning rate that is too high impedes precise updates of parameters, leading to sub-optimal results. On the other hand, an extremely low learning rate does not effectively learn all the necessary information within a restricted number of iterations, resulting in less than optimal performance. Notwithstanding these factors, our model exhibits strong performance throughout a broad spectrum of learning rates, offering adaptability for fine-tuning according to specific goals.

VII. DISCUSSION AND COMPARATIVE ANALYSIS

The simulation results presented in the previous section highlight the exceptional effectiveness of our proposed model in accurately differentiating cyberbullying from non-cyberbullying content. By utilizing cutting-edge deep learning methods, our model demonstrates outstanding efficiency and precision, even when handling large datasets. The training process is strong and produces a highly effective generation of modalities. One notable advantage is the skillful

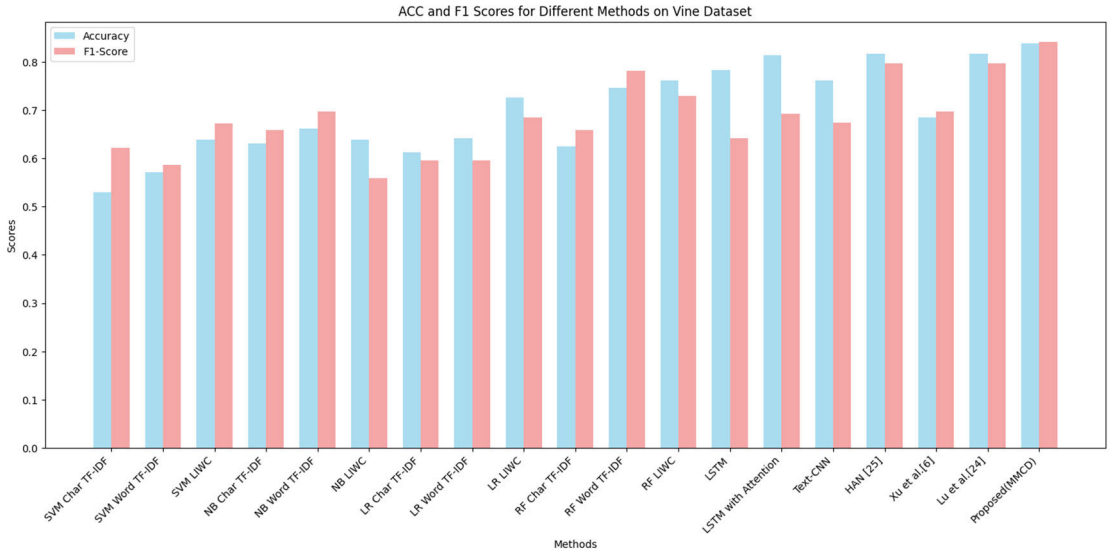

FIGURE 2. Accuracy and F1 scores for different methods on vine network dataset.

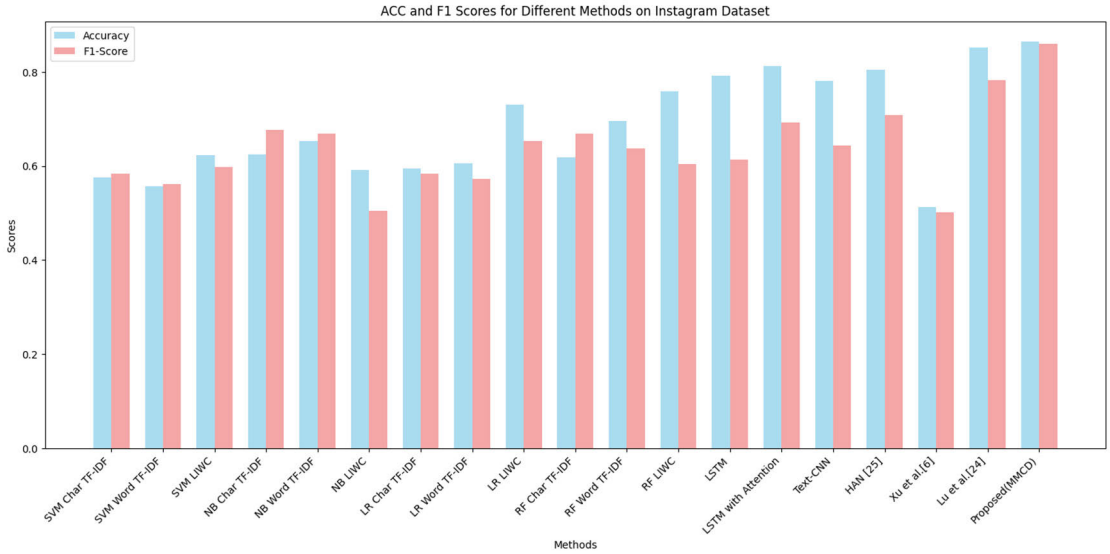

FIGURE 3. Accuracy and F1 scores for different methods on instagram network dataset.

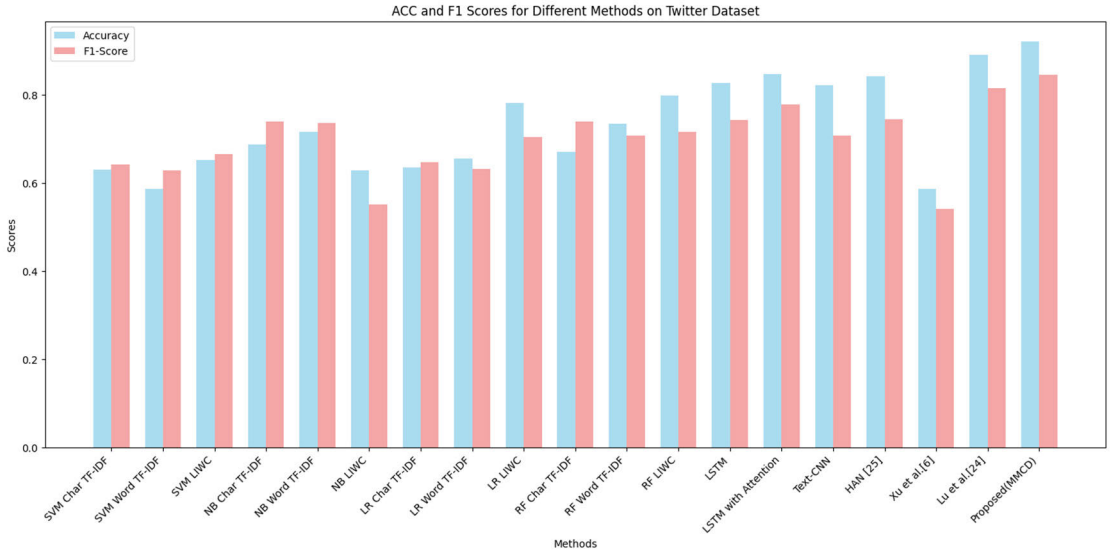

FIGURE 4. Accuracy and F1 scores for different methods on twitter network dataset.

integration of different modalities, which enhances overall performance. Our method differs from previous studies that only use one type of data. Instead, we combine four main types of data, text, comments, media, and metadata—which leads to a significant improvement in performance, as shown in Table 3. It is worth mentioning that models that use only two modalities consistently show worse results, which further emphasizes the superiority of our approach. Our model demonstrates better results compared to existing state-of-the-art models, as evidenced by higher accuracy and F-measure scores. The improved performance is observed across Twitter, Vine, and Instagram datasets. On the Twitter dataset, our model achieves an accuracy of 92.1% and an F-measure of 86.40%, surpassing other methods. Comparable levels of performance are consistently seen on Vine and Instagram datasets, with accuracy scores of 83.80% and 86.41%, respectively, along with corresponding F-measure scores. Our model on Instagram demonstrates an accuracy of 86.41% and an F-measure of 86%, highlighting its superiority compared to current cutting-edge models [34], [35], [36], [37]. The results highlight the consistent superiority of our method across different datasets and modalities, confirming its effectiveness in accurately identifying cyberbullying from non-cyberbullying content.

Moreover, the proposed framework utilizes distinct hyperparameters to enhance its performance. For example, the Twitter dataset uses a learning rate of 0.01, a batch size of 128, and an embedding dimension of 512. The selection of these hyperparameters aims to strike a balance between the complexity of the model and the efficiency of computational

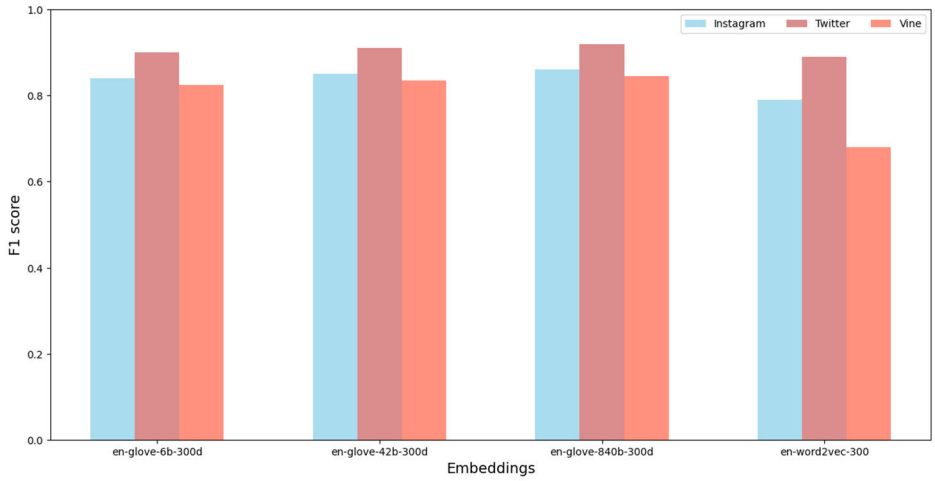

FIGURE 5. F1 scores across various word embeddings: an in-depth analysis on Twitter, Instagram and Vine datasets.

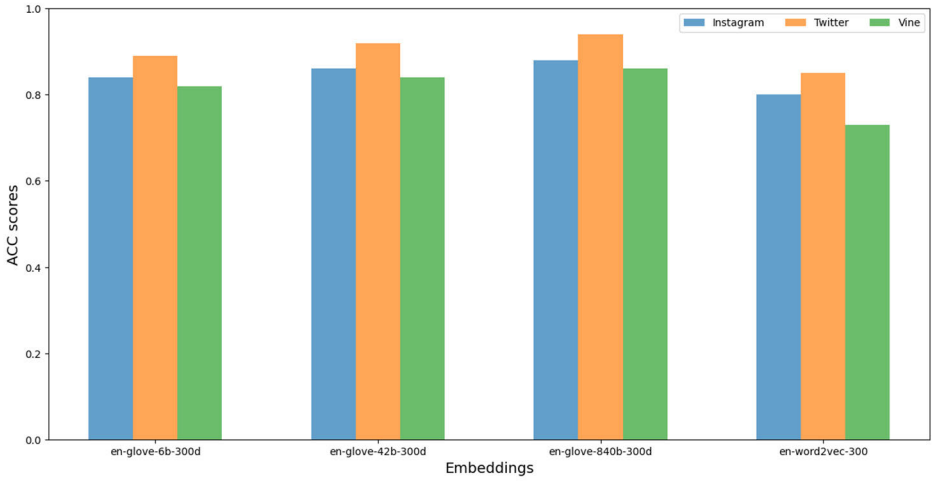

FIGURE 6. Accuracy scores across various word embeddings: an in-depth analysis on Twitter, Instagram and Vine datasets.

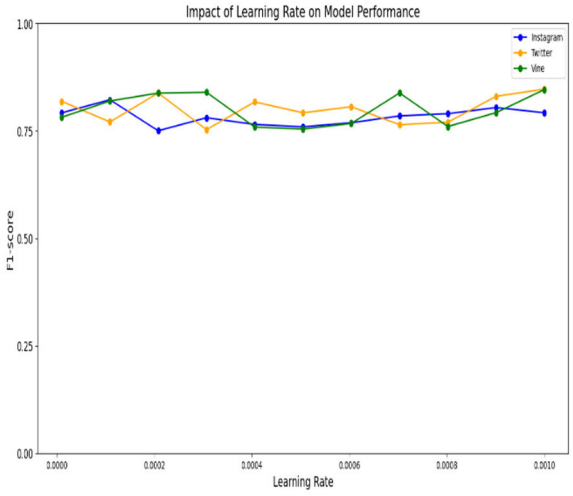

FIGURE 7. Exploring the impact of learning rate on model performance based on Twitter, Instagram, and Vine datasets.

TABLE 3. Comparison of state-of-art existing model and proposed model.

Studies		Approaches	Modalities	Performance Measure		
				Accuracy	Hyper-parameters	F1
27		2D-CNN, Inception V3, VGG-16	Graphical	89%	N/A	79.45 %
23		RCNN based Residual Bi LSTM	Textual and graphical	*N/A*	N/A	75%
2		Transformer, Bi-GRU, and CNN	Textual and Content	87.60 %	N/A	86.01 %
18		Attention based Bi-GRU	Textual and graphical	*N/A*	N/A	74%
Proposed	Twitter	Self-attention-based\\ BiLSTM model, Multi-Layer\\ Perceptron	Text, comments, media, and metadata can also support Visual contents, e.g., pictures, Memes	92.1%	LR. 0.01 Batch size *64, 128*	86.40 %
	Vin			83.80 %		84.10 %
	Instagram			86.41 %		86%

processes in order to achieve optimal training and inference. Similarly, the Vine dataset utilizes a learning rate of 0.001, a batch size of 64, and an embedding dimension of 300. These same hyperparameters are also used for the Instagram dataset. The meticulous selection of these hyperparameters enhances the model's capacity to efficiently acquire and apply patterns from various datasets, resulting in exceptional performance in detecting cyberbullying content.

Table 4 presents the model's performance on three separate datasets: Vine, Instagram, and Twitter. Each dataset contains both occurrences of cyberbullying and non-cyberbullying behavior. Within the Vine dataset, 85 instances of cyberbullying are correctly identified by the model out of a total of 100 instances. However, the model mistakenly classifies 15 instances as non-cyberbullying, leading to an error rate of 15%. The Twitter dataset, consisting of 10,500 instances of confirmed cyberbullying, is accurately identified by the model in 8,500 cases. However, the model incorrectly classifies 1,000 cases as non-cyberbullying, resulting in a 10% error rate. Regarding the Facebook dataset, the model accurately detects 12,000 occurrences of cyberbullying out of the total 15,000 confirmed instances. However, it incorrectly

TABLE 4. Confusion matrix of proposed model.

Dataset	True Non-CB	True CB	Predicted Non-CB	Predicted CB
Vine	666	304	616	354
Instagram	1540	678	1339	879
Twitter	15750	14250	15120	12380

classifies 3,000 instances as non-cyberbullying, leading to a 20% rate of misclassification.

The MMCD model's incapacity to integrate media data is a significant constraint that could have implications for the thoroughness of the outcomes. Media content, such as images and videos, frequently includes vital contextual information that can impact the understanding of text-based content. The MMCD model's omission of media data in the analysis may result in neglecting significant cues and subtleties that could enhance the accuracy of comprehending cyberbullying behavior. As a result, the accuracy of the model's predictions may be affected, resulting in incorrect categorizations or incomplete evaluations of cyberbullying cases. This constraint highlights the significance of thoroughly incorporating various data modalities in future versions of the MMCD model to guarantee a more comprehensive and nuanced approach to detecting cyberbullying.

The proposed model's ethical considerations revolve around privacy preservation, potential biases, and unintended consequences. Ensuring user privacy and data protection is paramount, especially when analyzing sensitive information from social media platforms. Additionally, mitigating biases in the model's training data and outputs is essential to prevent discriminatory outcomes or false accusations.

VIII. CONCLUSION

In this research, we introduce an innovative and reliable framework for detecting cyberbullying that utilizes three modules to extract unique information from different modalities inside a social network. The initial module utilizes bidirectional LSTM with attention techniques to effectively capture the intrinsic features of posts. To conduct a detailed study of post-comments, we propose the use of hierarchical attention networks. These networks operate dynamically at both the word and comment levels. Furthermore, to effectively encode meta-information, including video and image content, our methodology incorporates a Multilayer Perceptron. The sophisticated and flexible cyberbullying detection system that was created using this thorough methodology was put to the test by carefully analyzing three real datasets gathered from social platforms.

Subsequent investigations into the amalgamation of diverse information modalities for the purpose of cyberbullying detection are of abject importance. In the context of social media, this necessitates a meticulous analysis of the intricate interrelationships among diverse fields of information. An in-depth comprehension of growing trends and patterns is crucial for the modeling of emerging kinds of cyberbullying behavior. In addition, to improve the precision and usefulness of cyberbullying detection models, future endeavors should focus on improving current approaches, investigating sophisticated deep learning structures, and integrating real-time data processing capabilities. Moreover, doing a more in-depth examination of the socio-psychological factors related to cyberbullying and including explainable AI methods can enhance the creation of cyberbullying detection systems that are both more efficient and socially accountable.

REFERENCES

[1] S. Salawu, Y. He, and J. Lumsden, "Approaches to automated detection of cyberbullying: A survey," *IEEE Trans. Affect. Comput.*, vol. 11, no. 1, pp. 3–24, Jan. 2020.

[2] H. Dani, J. Li, and H. Liu, "Sentiment informed cyberbullying detection in social media," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, Skopje, Macedonia. Cham, Switzerland: Springer, Sep. 2017, pp. 52–67.

[3] A. Mangaonkar, A. Hayrapetian, and R. Raje, "Collaborative detection of cyberbullying behavior in Twitter data," in *Proc. IEEE Int. Conf. Electro/Information Technol. (EIT)*, May 2015, pp. 611–616.

[4] Z. Li, J. Kawamoto, Y. Feng, and K. Sakurai, "Cyberbullying detection using parent-child relationship between comments," in *Proc. 18th Int. Conf. Inf. Integr. Web-Based Appl. Services*, Nov. 2016, pp. 325–334.

[5] D. Soni and V. K. Singh, "See no evil, hear no evil: Audio-visual-textual cyberbullying detection," *Proc. ACM Human-Comput. Interact.*, vol. 2, no. CSCW, pp. 1–26, Nov. 2018.

[6] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2016, pp. 1480–1489.

[7] P. Burnap and M. L. Williams, "Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making," *Policy Internet*, vol. 7, no. 2, pp. 223–242, Jun. 2015.

[8] M. Dadvar, D. Trieschnigg, and F. De Jong, "Experts and machines against bullies: A hybrid approach to detect cyberbullies," in *Proc. 27th Can. Conf. Artif. Intell.*, Montréal, QC, Canada: Springer, May 2014, pp. 275–281.

[9] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," in *Proc. 17th Int. Conf. Distrib. Comput. Netw.*, Jan. 2016, pp. 1–6.

[10] P. Burnap and M. L. Williams, "Us and them: Identifying cyber hate on Twitter across multiple protected characteristics," *EPJ Data Sci.*, vol. 5, no. 1, pp. 1–15, Dec. 2016.

[11] L. Cheng, J. Li, Y. N. Silva, D. L. Hall, and H. Liu, "XBully: Cyberbullying detection within a multi-modal context," in *Proc. 12th ACM Int. Conf. Web Search Data Mining*, Jan. 2019, pp. 339–347.

[12] Y. Li and J. Ye, "Learning adversarial networks for semi-supervised text classification via policy gradient," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 1715–1723.

[13] V. S. Chavan and S. Shylaja, "Machine learning approach for detection of cyber-aggressive comments by peers on social media network," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Aug. 2015, pp. 2354–2467.

[14] M. A. Al-Garadi, K. D. Varathan, and S. D. Ravana, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network," *Comput. Hum. Behav.*, vol. 63, pp. 433–443, Oct. 2016.

[15] C. Chelmis, D.-S. Zois, and M. Yao, "Mining patterns of cyberbullying on Twitter," in *Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2017, pp. 126–133.

[16] L. Xu, Z. Wang, S. Zhang, X. Yuan, M. Wang, and E. Chen, "Modeling student performance using feature crosses information for knowledge tracing," *IEEE Trans. Learn. Technol.*, vol. 8, no. 4, pp. 1–14, Mar. 2024.

[17] B. Gambäck and U. K. Sikdar, "Using convolutional neural networks to classify hate-speech," in *Proc. 1st Workshop Abusive Lang. Online*, 2017, pp. 85–90.

[18] H. Hosseinmardi, S. Li, Z. Yang, Q. Lv, R. I. Rafiq, R. Han, and S. Mishra, "A comparison of common users across Instagram and Ask.fm to better understand cyberbullying," in *Proc. IEEE 4th Int. Conf. Big Data Cloud Comput.*, Dec. 2014, pp. 355–362.
[19] R. Zhang, H. Lee, and D. Radev, "Dependency sensitive convolutional neural networks for modeling sentences and documents," 2016, *arXiv:1611.02361*.
[20] J. H. Park and P. Fung, "One-step and two-step classification for abusive language detection on Twitter," 2017, *arXiv:1706.01206*.
[21] Z. Zhang, D. Robinson, and J. Tepper, "Detecting hate speech on Twitter using a convolution-GRU based deep neural network," in *Proc. Eur. Semantic Web Conf.* Cham, Switzerland: Springer, 2018, pp. 745–760.
[22] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu, "Attention-based bidirectional long short-term memory networks for relation classification," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 207–212.
[23] A. Zhang, B. Li, S. Wan, and K. Wang, "Cyberbullying detection with BiRNN and attention mechanism," in *Proc. Int. Conf. Mach. Learn. Intell. Commun.* Cham, Switzerland: Springer, Jul. 2019, pp. 623–635.
[24] A. M. Founta, D. Chatzakou, N. Kourtellis, J. Blackburn, A. Vakali, and I. Leontiadis, "A unified deep learning architecture for abuse detection," in *Proc. 10th ACM Conf. Web Sci.*, Jun. 2019, pp. 105–114.
[25] W. M. S. Yafooz, A. Al-Dhaqm, and A. Alsaeedi, "Detecting kids cyberbullying using transfer learning approach: Transformer fine-tuning models," in *Kids Cybersecurity Using Computational Intelligence Techniques*. Cham, Switzerland: Springer, 2023, pp. 255–267.
[26] R. Alhejaili, A. Alsaeedi, and W. M. Yafooz, "Detecting hate speech in Arabic tweets during COVID-19 using machine learning approaches," in *Proc. 3rd Doctoral Symp. Comput. Intell. (DoSCI)*. Singapore: Springer, Nov. 2022, pp. 467–475.
[27] L. Cheng, R. Guo, Y. Silva, D. Hall, and H. Liu, "Hierarchical attention networks for cyberbullying detection on the Instagram social network," in *Proc. SIAM Int. Conf. Data Mining*, 2019, pp. 235–243.
[28] R. Ibn Rafiq, H. Hosseinmardi, R. Han, Q. Lv, S. Mishra, and S. A. Mattson, "Careful what you share in six seconds: Detecting cyberbullying instances in vine," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2015, pp. 617–622.
[29] H. Hosseinmardi, S. A. Mattson, R. Ibn Rafiq, R. Han, Q. Lv, and S. Mishra, "Analyzing labeled cyberbullying incidents on the Instagram social network," in *Proc. 7th Int. Conf. SocInfo*, Beijing, China: Springer, Dec. 2015, pp. 49–66.
[30] J. O. Atoum, "Detecting cyberbullying from tweets through machine learning techniques with sentiment analysis," in *Proc. Future Inf. Commun. Conf.* Cham, Switzerland: Springer, Mar. 2023, pp. 25–38.
[31] I. S. Ahmad, M. F. Darmawan, and C. A. Talib, "Cyberbullying awareness through sentiment analysis based on Twitter," in *Kids Cybersecurity Using Computational Intelligence Techniques*. Cham, Switzerland: Springer, 2023, pp. 195–211.
[32] S. I. Alqahtani, W. M. S. Yafooz, A. Alsaeedi, L. Syed, and R. Alluhaibi, "Children's safety on YouTube: A systematic review," *Appl. Sci.*, vol. 13, no. 6, p. 4044, Mar. 2023.
[33] L. Wang and T. Islam, "Automatic detection of cyberbullying: Racism and sexism on Twitter," in *Proc. 14th Int. Conf. Global Security, Safety Sustainability*, London, U.K. Cham, Switzerland: Springer, Jan. 2023, pp. 105–122.
[34] P. K. Roy and F. U. Mali, "Cyberbullying detection using deep transfer learning," *Complex Intell. Syst.*, vol. 8, no. 6, pp. 5449–5467, Dec. 2022.
[35] I. Musyoka, J. Wandeto, and B. Kituku, "Multimodal cyberbullying detection using deep learning techniques: A review," in *Proc. Int. Conf. Inf. Commun. Technol. Develop. Afr. (ICT4DA)*, Oct. 2023, pp. 187–192.
[36] M. Alotaibi, B. Alotaibi, and A. Razaque, "A multichannel deep learning framework for cyberbullying detection on social media," *Electronics*, vol. 10, no. 21, p. 2664, Oct. 2021.
[37] K. Kumari, J. P. Singh, Y. K. Dwivedi, and N. P. Rana, "Multi-modal aggression identification using convolutional neural network and binary particle swarm optimization," *Future Gener. Comput. Syst.*, vol. 118, pp. 187–197, May 2021.

MAHMOUD AHMAD AL-KHASAWNEH received the B.Sc. degree in computer science from Yarmouk University, Jordan, in 2003, and the M.Sc. and Ph.D. degrees in computer science from Universiti Teknologi Malaysia (UTM), Johor, Malaysia, in 2013 and 2018, respectively. He is currently an Assistant Professor with the School of Computing, Skyline University College, University City Sharjah, Sharjah, United Arab Emirates. His research interests include security, image encryption, wireless networks, blockchain, machine learning, the Internet of Things, and big data.

MUHAMMAD FAHEEM (Member, IEEE) received the B.Sc. degree in computer engineering from the Department of Computer Engineering, University College of Engineering and Technology, Bahauddin Zakariya University, Multan, Pakistan, in 2010, the M.S. degree in computer science from the Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, in 2012, and the Ph.D. degree in computer science from the Faculty of Engineering, Universiti Teknologi Malaysia, Johor Bahru, Malaysia, in 2021. Previously, he was a Lecturer with the COMSATS Institute of Information and Technology, Pakistan, from 2012 to 2014. He was an Assistant Professor with the Department of Computer Engineering, Abdullah Gul University, Turkey, from 2014 to 2022. Currently, he is a Senior Researcher with the School of Computing (Innovations and Technology), University of Vaasa, Vaasa, Finland. He has authored several papers in refereed journals and conferences and served as a reviewer for numerous journals in IEEE, Elsevier, Springer, Willey, Hindawi, and MDPI. His research interests include cybersecurity, blockchain, smart grids, smart cities, and industry 4.0.

ALA ABDULSALAM ALAROOD received the bachelor's and master's degrees in computer science from Yarmok University, Jordan, and the Ph.D. degree in computer science from the University of Technology Malaysia. He is currently an Assistant Professor of computer science with the Faculty of Computer and Information Technology, University of Jeddah. His current research interests include information security, network security, steganalysis, machine learning, and neural networks.

SAFA HABIBULLAH is currently an Assistant Professor of computer science with the Faculty of Computer and Information Technology, University of Jeddah. Her current research interests include security, machine learning, and neural networks.

EESA ALSOLAMI received the bachelor's degree in computer science from King Abdulaziz University, Saudi Arabia, in 2002, and the master's degree in information technology and the Ph.D. degree in information security from the Queensland University of Technology, Australia, in 2008 and 2012, respectively. He is currently an Assistant Professor with the Department of Information Technology, University of Jeddah, Saudi Arabia. His research interests include information security and biometric technology.