This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3504258

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Enhancing Immoral Post Detection on Social Networks Using Knowledge Graph and Attention-Based BiLSTM Framework

Bibi Saqia[1], Khairullah Khan[1], Atta Ur Rahman[2], Sajid Ullah Khan[3], Mohammed Alkhowaiter[3], Wahab Khan[1], Ashraf Ullah[1]

[1]Department of Computer Science, University of Science and Technology Bannu, 28100, Pakistan

[2]Riphah Institute of System Engineering (RISE), Riphah International University Islamabad, 46000, Pakistan

[3]Department of Computer Engineering, College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, Alharj, KSA

* Correspondence authors: Bibi Saqia (saqiaktk@ustb.edu.pk) and Sajid Ullah Khan (sk.khan@psau.edu.sa)

ABSTRACT Preserving a secure and morally safe online environment on social media is a challenging task. It is essential to find immoral or unsuitable information in user-generated postings to safeguard users and enforce community standards. Various Natural Language Processing (NLP) approaches are being employed to detect subtle immoral posts; however, there remains a research gap due to the semantic and contextual complexity of natural language. To bridge this gap, this work proposes the use of a Knowledge Graph (KG) for entity recognition and the extraction of semantic relationships in Social Network (SN) posts. By doing so, the KG helps provide a deeper contextual understanding, enabling the detection of negative interactions between entities that are often present in immoral content. KG allows us to extract these associations from the text, enabling the model to recognize language that leads to immoral behavior. By utilizing a KG, the model can more easily identify connections between entities, verify statements made in postings, and classify material more precisely. GloVe (Global Vector) word embedding is used to transform the enriched text data into numerical representations. An attention-based Bidirectional Long Short-Term Memory (BiLSTM) network performs the classification task. The BiLSTM concurrently analyses the input sequence in both directions, enabling the network to recognize not only the context that is present at the moment but also the context in which each word in the sequence comes before and after. To validate the model performance, we used benchmark datasets Self-Annotated Reddit Corpus (SARC), and Hate Evaluation (HatEval) dataset. We achieved a higher F1-score of 82.79% and 84.06% on both datasets and outperformed state-of-the-art works.

INDEX TERMS Immoral posts, Post detection, NLP, Social media, Knowledge graph, BiLSTM network

I. INTRODUCTION

SN provides global interaction, sharing, and self-expression through modern communication. On the other hand, this extensive accessibility has also led to a rise in offensive posts on these sites [1, 2]. The offensive content might include Hate Speech (HS), profanity, discriminatory remarks, and cyberbullying [3]. Previous studies applied traditional machine learning (ML) approaches for various social media text classification tasks [4]. However, these methods often face challenges, such as trouble in handling large and different datasets, issues in feature extraction, and a lack of flexibility in changing language patterns in SN. Consequently, these traditional approaches may not be as successful in identifying nuanced and context-dependent content, such as unethical posts [5]. In text mining, the performance of ML approaches is considerably influenced by the manual feature extraction procedure [6-8]. Domain experts, permitting the integration of domain-specific knowledge that can increase the model's productivity, often make manually designed features. However, these methods are time-consuming, require more effort, and lack in interpreting the contextual meaning between words [9]. Understanding texts usually requires taking into account the larger context in which it is delivered [10]. Traditional text-based ML techniques did not adequately extract the context, resulting in a limited and imperfect understanding [11, 12]. These methods typically rely on statistical patterns discovered via the analysis of big datasets [13]. The consistency of the responses may decline if the text's subject matter extends beyond the system's training data. NLP has many different phrase forms and grammatical patterns, making it potentially complex. It is inherently ambiguous; depending on the context, a single word or phrase can convey different meanings. Analyzing such a wide variety of linguistic constructs is challenging for traditional ML methods. Consequently, it is essential to create automatic and intelligent tools for identifying hate, offensive, and immoral posts on SN [14]. Social media slang is casual language, words, and phrases used on digital networks. It illustrates how internet communication is dynamic. Researchers can gain important insights by examining social media terminology. It makes it possible to comprehend sentiment analysis, communication patterns, and online user behavior more deeply [15]. The proposed work specifically focused on sarcasm, hate speech, and cyberbullying prediction on SN. These types of content are prominent examples of harmful and immoral behavior on social media. While our primary focus is on these categories, they are all considered forms of immoral content. Consequently, our model is designed to detect a broad range of harmful behaviors, acknowledging that these types of content often overlap and contribute to the overall toxicity in online environments. Immoral content generally refers to posts promoting hate speech, discrimination, violence, or actions that violate widely accepted ethical standards, as outlined by social media platforms and regulatory guidelines [16]. Thus, an immoral post breaches commonly recognized ethical norms and promotes harm, hatred, bias, violence, or exploitation. Moral posts, on the other hand, follow moral standards, encourage civil conversation, and benefit the community. Table 1 represents the features of moral and immoral posts to highlight the dissimilarities in ethical and unethical conduct across social networks. Content moderation plays a crucial role in limiting the spread of such content, ensuring a safer online environment for users. Recent studies have integrated Large Language Models (LLMs), such as ChatGPT, to enhance the personalization and fairness of moderation systems, making them more adaptable and user-friendly [17]. However, the definition of acceptable content varies significantly across cultures and individuals, complicating the application of universal moderation guidelines. A balance is needed between removing harmful content and respecting freedom of expression, but current one-size-fits-all approaches often fail to account for individual preferences. Personalization in content moderation, where users' thresholds for what they consider immoral or offensive can differ, has emerged as an area of focus, offering insights into how personalized moderation can improve user engagement and transparency [18].

The immoral posts detection of SN has been extensively investigated by academics using a variety of methods. These methods frequently make use of sentiment analysis, ML, and DL approaches. From the literature, we observed that the integration of intelligent feature extraction along with an advanced DL model improves the accuracy of immoral content detection. Therefore, this work deployed the integration of KG-based BiLSTM networks to improve the recognition of the immoral text on SN. The proposed model efficiently handles the ever-changing nature of unethical content on social media networks through a KG that is updated with new data over time. This continuous updating assists the model in keeping up with evolving trends and changes in linguistics. Furthermore, the Attention-Based BiLSTM framework certifies that the model can emphasize the most significant components of the content, which is important for identifying new types of immoral conduct. Frequently integrating new training data into the model assists in keeping its accuracy and sensitivity to changing social media content. The following are the main contributions of this work:

TABLE 1. MORAL VS. IMMORAL POSTS: INTENT, BEHAVIOR, AND IMPACT

Category	Moral Post	Immoral Post
Definition	Content that aligns with moral norms and encourages positive values.	Content that violates moral standards and stimulates damage, discrimination, or offensive conduct.
Instances	Positive posts, empathy educational content, encouragement,	Cyberbullying, hate speech, offensive language, threats, and extremism.
Tone	Respectful, inclusive, constructive.	Disrespectful, discriminatory, hurtful.
Language Used	Neutral or positive language; polite, kind, non-violent.	Offensive language, slurs, harmful stereotypes, threats.
Intent	To inform, educate, or inspire.	To harm, insult, intimidate, or deceive.
Impact on people	Encourages understanding, support, and collaboration.	Leads to conflict, distress, division, or harm.
Legal Implications	Usually obeys with laws and community guidelines.	May violate laws (e.g., hate speech laws) and platform policies.

1. This work presents a novel KG-based feature extraction approach for immoral post-detection on social media. The adoption of KG enables the understanding of complicated interactions between words in social media posts, which results in more accurate detection.
2. To extract the contextual meaning between words, this work deploys an attention-based BiLSTM network. It

processes the words in both forward and backward directions. It enables it to recognize the context from both past and future tokens.

3. The proposed model combines KG and BiLSTM, to correctly differentiate between moral and immoral content on SN.

The remaining paper is organized as follows: Section 2, describes the literature review. Section 3 explains the proposed model. Section 4 discusses the experiment conducted in this work. Section 5 presents the results and discussion. Finally, section 6 concludes the study with future direction.

II.BACKGROUND

Numerous studies have been conducted recently on social media content analysis and immoral post detection [19]. An immoral post is any online material that disobeys moral and ethical norms and frequently encourages violence, hatred, and other unethical behavior [20]. The work published by [21] presents a novel multilingual hybrid dataset merging monolingual and bilingual resources. They used a dataset of 42,560 social media comments. They obtained an average F1 score of 0.79 for monolingual tasks and 0.86 for code-mixed and script-mixed tasks. Their results improve offensive language detection methods and report the issues of multilingual social media platforms. The study performed in [22] provided a dataset specially designed for identifying offensive text in the Algerian language. The dataset comprises 14,150 annotated texts from YouTube, Twitter, and Facebook. They deploy various ML and DL models on the collected dataset. Among these, the Bi-GRU approach obtained the highest F1-score of 75.8% and an accuracy of 73.6% for immoral post detection. The research published in [23] focused on text attributes for abusive linguistic detection. The user view is characterized using the user's previous history of attitude and ethics. They developed a dataset with suitable users and tweets circulation to decrease the imbalance issues of distribution of users and tweets which could affect the generalization capability of their model. Their proposed technique achieved an F1 score of 89.94%. The work conducted in [24] discovered word-level features such as N-gram, and PoS tags to identify irony and critical remarks across three benchmark datasets. Both manual and word embedding attributes were compared with ML and hybrid deep learning models. Their proposed technique attained 87% accuracy with manual attributes in deep learning techniques and 92% by Random Forest technique. The work published in [25] presented an innovative method where syntax dependency graphs were created for each occurrence by syntactical data. They used the emotion transfer scheme and dependency graph convolutional, to improve contextual learning. Furthermore, opinion resources that are similar to the pieces of information were combined as a secondary job to share related opinion features and increase identification accuracy. Their proposed technique outperforms current approaches, attaining a 3.88% and 0.54% enhancement in Macro-F1 scores on the OLID and HatEval datasets, correspondingly. Different researchers and interested field workers work together to create computer models. Which can quickly and accurately identify and categorize objectionable content and put a stop to it to preserve the social media network. Consequently, the past several years have seen a remarkable advancement in research on offensive language detection [26, 27]. Table 2 shows a summary of previous related research.

A. LIMITATION OF PREVIOUS WORK

Due to the high throughput and volume of content produced on SN, it becomes difficult to identify immoral posts. To stop HS from spreading, it is essential to identify it quickly [28]. Identifying the massive content of data circulating on social media environments is a challenging task [4]. Previous work employed traditional approaches that frequently rely on manual feature extraction, which can be laborious and time-intensive and may not be well suited to the variety and evolution of immoral posts [29]. Previous techniques are unable to identify immoral content from the dynamic social media content. It is difficult to successfully detect subtle, context-dependent signs of immoral posts. Because of the complexity of several languages and cultural nuances included in the social media text, these methodologies may produce biased results [30, 31]. Prior studies emphasized a narrow set of features, such as basic text or keyword-based features, without integrating further advanced language or background components. This limitation can obstruct the model's capacity to identify delicate forms of immoral content, such as hate speech or sarcasm [32]. Previous methods did not adapt extensive knowledge sources, like KG, which can improve the understanding of the associations between the objects and context of the posts. Social media data and trends change quickly, and manual feature extraction techniques cannot adapt well to these variations [33].

The work published in [34] explored analyses of privacy among people of the Arab Gulf. They studied how privacy was accomplished and assumed in technology-mediated atmospheres, emphasizing the substantial effect of Islam and cultural civilizations in determining privacy standards. Their work also accessible socially sensitive strategy values and suggests future work to integrate earlier unexplored features of privacy, which inspire how people are involved with social platforms. Their study provides appreciated perceptions of the cultural issues affecting privacy observed in the Arab Gulf.

The work performed by [35] discussed the harmless and trustable usage of social networks for vulnerable women, mostly those influenced by Arab culture and Islam. They concentrated on people such as conventional Sunni Muslims, containing Syrian war immigrants, who face increased susceptibility due to issues such as lack of digital literacy,

TABLE 2. SUMMARY OF PREVIOUS RESEARCH

Study	Year	SN	Dataset	Approach	Main Innovation	Results			
						P	R	F1-score	Accuracy
d'Sa, Illina [36]	2020	Twitter	24,883 tweets	BERT Model	Detection of toxic speech, word embedding-based classification	------	------	91.9	---------
Sharma, Singh [37]	2022	Reddit	Self-Annotated Reddit Corpus (SARC)	Auto-Encoder Model	Development of a hybrid model for sarcasm detection	0.83	0.85	0.84	83.92
Song, Giunchiglia [10]	2022	Twitter	HatEval Dataset	RSGNN	Offensive language detection on social media, Post content, user interactions, Improved detection accuracy	74.29	74.14	74.21	74.04
Mossie and Wang [38]	2018	Facebook	Amharic linguistic across Facebook	RF and NB	Formation of a distributed model for Amharic hate speech identification using Apache Spark.	------	------	------	79.83
Sap, Card [39]	2019	Twitter	Toxic Language Datasets	Dialect, Race Priming	To decrease annotator bias in toxic linguistic detection.	------	------	------	84.1
Kumar, Naraparedd y [40]	2020	Reddit	SARC dataset	SVM, MHA-BiLSTM	Handcrafted Features, SVM Model, Multi-Head Attention BiLSTM (MHA-BiLSTM), Sarcasm Detection, Performance Enhancement .	60.26	53.71	------	56.79
Struß, Siegel [41]	2019	Twitter	Twitter dataset	Coarse-grained binary multi-class	Classification task for abusive texts, separation of explicit from	76.72	76.84	76.78	----------

					implicit offensive tweets, augmentation of the dataset,				
Romim, Ahmed [42]	2021	Facebook	HS-BAN dataset covers 50,000 Bangla remarks	Bi-LSTM model	Formation of HS-BAN, Decreasing annotation bias including slang differences.	------	------	86.78	-----------
Kumar, Ojha [43]	2018	Facebook, Twitter	15,000 Facebook remarks in Hindi and English	SVM, Neural Networks	Aggression Identification Challenge, Neural Networks vs. Traditional Models, Annotation Consistency.	------	------	0.64	----------
Raza, Memon [44]	2020	Online Forums	Cyberbullying dataset	Supervised ML	Ensemble ML techniques, Voting Classifier, Cyberbullying Detection	------	-----	------	84.4
Subramanian, Ponnusamy [45]	2022	YouTube	YouTube remarks in Tamil	XLM-RoBERT	RoBERTa Model, Aggressive Text Detection, Code-Mixed Tamil Text	67.79	68.37	68.08	88.53

insufficient resources, and cultural biases. Their study recognized critical factors that communication systems should integrate to better help these helpless people. By employing ongoing investigation, they suggested technological methods to identify these worries and presented the practicability of one of these features by adding it to a traditional communication system. This work conveyed significant visions into how social networks can be modified to improve secure vulnerable handlers in socially sensitive circumstances.

Some earlier techniques may have had trouble capturing sophisticated contextual information, making it impossible to tell the difference between content that is immoral and stuff that is merely harmless but might seem the same when taken out of context [46]. This may result often traditional methods not efficiently handling to capture immoral content.

By addressing these critical aspects, the proposed work tries to overcome these limitations and describes a more robust and accurate method for identifying immoral posts on social platforms. We emphasized the significance of automated methods for noticing immoral posts and the possible merits of including KG. Utilizing KG-based feature extraction in conjunction with BiLSTM networks improves the process of immoral post detection.

III. METHODOLOGY

This section discusses the entire process of the immoral posts' detection from social media platforms. The proposed model deals with the problem of identifying immoral posts on SN sites, which frequently entails subtle and context-dependent concerns. The integration of KG made it easier to extract key features and capture semantic relationships between various posts. However, we encountered various challenges in integrating KG into the proposed model. Handling the vast and diverse nature of social media content using KG was challenging. Slang, irony, and other nonverbal phrases are used in social media texts. Correct interpretation and encoding of this data into a KG become challenging. The nature of morality is subjective. It differs among cultures, geographical areas, and even people. It is an intricate process to encode such an abstract and complicated idea into an organized knowledge network. This required us to carefully select and update relevant entities and relationships to ensure the graph accurately reflected the evolving language and trends. Integrating various sources of data and maintaining the graph's scalability as new information emerged were carefully handled. The proposed BiLSTM networks are fine-tuned to increase their efficiency for the task of immoral post detection. The initial step is to fine-tune the BiLSTM networks' hyperparameters. To prevent overfitting, a dynamic learning rate schedule is added to the training process. The data augmentation approaches are used to improve the model's generalization abilities. Figure 1 represents the block diagram of the proposed study.

A. DATASET

The SARC [37], and HatEval [10] benchmark datasets include a wide variety of user-generated posts from several social media sites. The posts address many types of immoral posts, including sarcastic, HS, rude language, and discriminating statements. Each post has a label indicating whether it is moral (positive class), immoral (negative class) sarcastic, or non-sarcastic. To maintain consistency, we make a detailed preprocessing step. The data is cleaned up, tokenized, and converted to lowercase. We eliminated frequent stop words, took lemmatization into account for dimensionality reduction, and addressed problems with data imbalance. Sequences have been cleaned to ensure uniform length, and embeddings utilized to build input features that incorporated semantic data. Target labels are converted into numerical quantities by label encoding. We prepared the datasets for model training and evaluation by properly separating and loading the data. This thorough preprocessing makes sure that the data is consistent and appropriate for the purpose study.

B. KG CONSTRUCTION

The KG developed an entire data structure that combines attributes, nodes, edges, and semantic information [47]. This organized knowledge is used by ML algorithms to efficiently identify immoral behavior and its context [48]. The building of a KG can be described using triples, where every triple contains a subject, a predicate, and an object [49]. The KG's fundamental structure is denoted by (s, p, and o). The subject entity is represented by s. While p stands for the predicate, which denotes how the subject and object are related. The object entity is represented by o. These triples are arranged into a graph structure, where entities are nodes and associations are edges, to create the KG. To enhance immoral posts detection techniques, the graph enables effective traversal, pattern identification, and contextual comprehension of social media posts. In this work, we utilize nodes to represent various entities, including users and posts, within the proposed network model. These nodes are interconnected by edges, which denote relationships or connections between the entities. For example, a user node may be connected to a post node through an edge to signify that the user has created or interacted with the post. Creating a KG entails expressing connections between things. The KG serves as a thorough semantic representation that improves comprehension of the intricate connections found in social media data. We assume, the simple KG as (Khan, "works at", Company B). Where, Khan is the subject here, while "works at" is the predicate denoting the connection. Company B is the object entity. This triple reveals Khan's employment with Company B. The KG's basis is made up of these triples as a whole, which capture relationships and semantic data systematically. Figure 2 describes the relationships between people and their postings in the context of detecting moral and immoral material on SN by KG construction. The numbers in Figure 2 show the similarity between different posts. Whereas U represents User while P represents Posts. KG is designed to show the similarity between different users and their posts. KG calculates the similarity in number using the content of each post. The KG has three primary entities: Users (U_1, U_2, U_3, U_4, U_5), Posts (P_1, P_2, P_3, P_4, P_5), and content types (Moral and Immoral). The 'authored' connection connects each user entity to the posts they created.

For example, U_1 wrote P_1, U_2 wrote P_2, and so on. This connection allows us to correlate each post with its author. To build the semantic aspect of the knowledge network, we proposed the 'conveys meaning of' connection. Posts are linked to the appropriate content types. P_1 conveys the meaning of morals, P_2 conveys the meaning of immoral, and so on. This connection allows us to categorize postings as having moral or immoral content.

We used contextual and semantic information inherent in user, post, and content type connections by using KG. This enables us to understand and investigate the complex network of relationships between users and their postings from the perspective of morality. It gave vital insights into user characteristics and the nature of their material, resulting in more effective and context-aware immoral post detection algorithms. Table 3 defines the example sentences of KG representation.

1) FEATURE EXTRACTION FROM THE KG

After the KG construction, we derive meaningful feature representations from the KG. Node, Edge, and Graph embedding's is a multi-step process of feature extraction used in this work. Assume, that G represents the KG with nodes and edges. Each edge in the graph is connected to attributes and each node is connected to a set of attributes.

Node Embedding's: Graph embedding techniques are used to embed graph nodes into a continuous vector space as defined in Eq.1.

$$N(v) = \int \left(A(v)\right) \tag{1}$$

Whereas, N (v) is a representation of the node's embedding created by applying the function f to the node's attributes A (v).

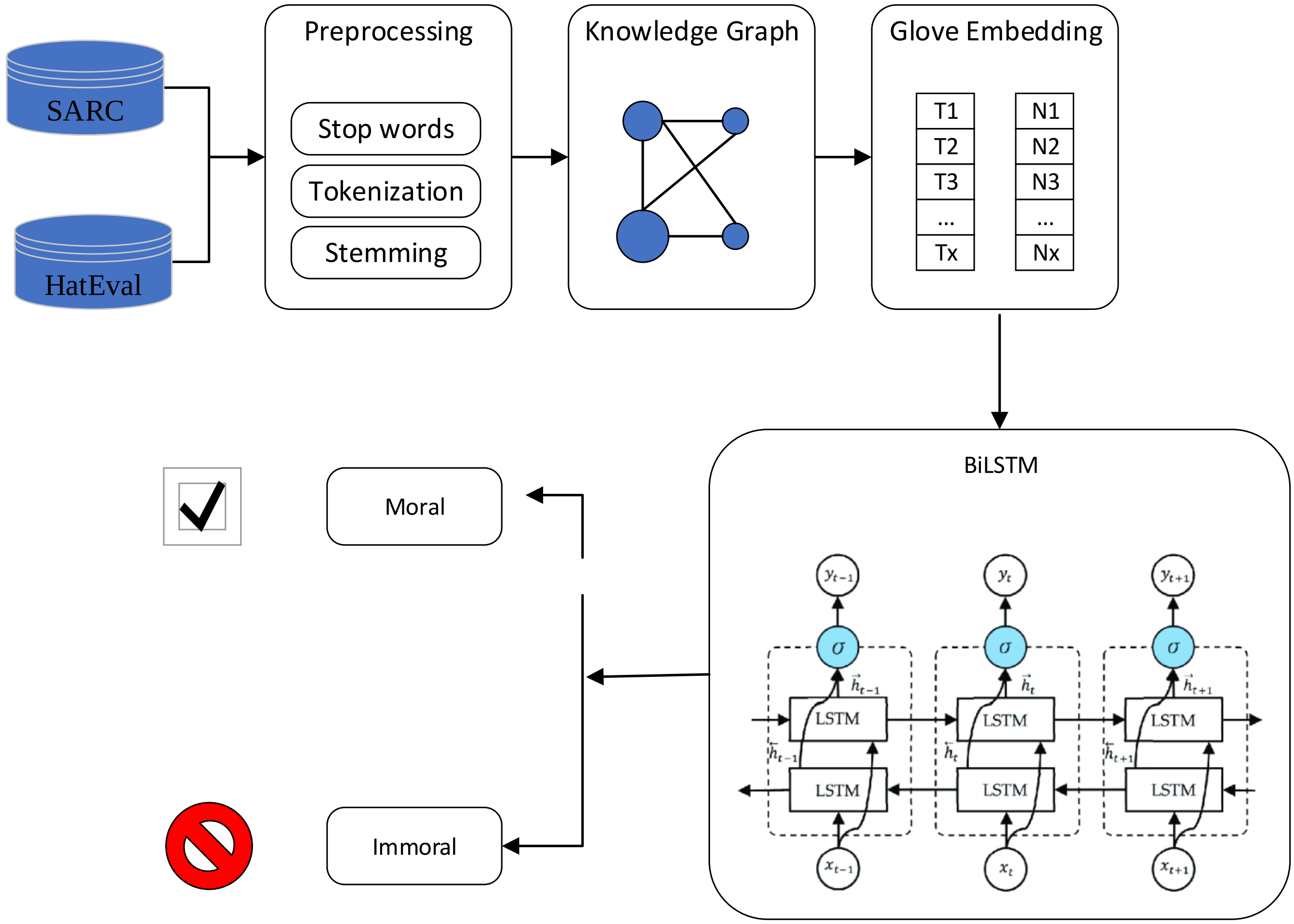

FIGURE 1. BLOCK DIAGRAM OF THE PROPOSED STUDY

Edge Embeddings: Edges can also be embedded using characteristics, just like nodes. Eq.2 indicates the edge embedding function.

$$E(e) = \int \left(A(e)\right) \qquad (2)$$

E (e) is the edge's embedding representation, which is obtained by applying a function ∫ to the edge's attributes A (e) as defined in Eq.3.

$$G(G) = g(\{N(v)|v \in V\},\{E(e)|e \in E\} \qquad (3)$$

Where G (G) denotes the function g-based aggregated embedding of the complete graph G. The equations effectively express how node and edge attributes in a KG are used to construct embeddings.

Word Embedding: The word embeddings is employed to capture a word's semantic meaning that depends on its context usage [50]. These methods seek to automatically learn concise and insightful textual representations, allowing for the accurate classification of immoral posts. The representation of words in an incessant vector space, which captures semantic links, makes word embeddings crucial in problems involving NLP. Let d be the dimension of the word embeddings and V be the vocabulary that contains V distinct words.

As Eq.4 states initially, words are denoted as one-hot vectors with only the word's index set to 1 and all other indices set to 0.

$$OneHot(w) = [0,0, \dots ,0] \qquad (4)$$

The word to be encoded is represented by w. Eq.5 signifies the encoding process during training, the word embedding matrix E∈R|V|×d is learned.

$$E = \begin{bmatrix} \vec{w}_1 \\ \vec{w}_2 \\ \vdots \\ \vec{w}_{|V|} \end{bmatrix} \qquad (5)$$

The word embedding vector for the ith word in the lexicon is represented by $\vec{w}$. The matrix-vector multiplication between

TABLE 3. EXAMPLE SENTENCES OF KG REPRESENTATION

Type	Example	KG representation
Immoral	It is wrong to promote hate speech and discrimination online.	(Concept: Hate Speech, Relationship: Promote, Concept: Wrong)
Moral	Giving to a charity to assist those in need is a moral deed.	(Concept: Charity, Relationship: Assist, Concept: Moral deed)
Immoral	Stealing someone's property is unethical and illegal. Property theft is immoral and unlawful.	(Concept: Theft, Relationship: Immoral, Concept: Unlawful)
Moral	It's kind to help an elderly neighbor with their groceries.	(Concept: Kindness, Relationship: Help, Concept: Good)
Immoral	Cheating on examinations demoralizes the educational system.	(Concept: Cheating, Relationship: Demoralizes, Concept: Educational System)
Moral	Being compassionate is saving and helping an injured animal.	(Concept: Saving, Relationship: Provides, Concept: Compassionate)

TABLE 4. EXAMPLE SENTENCE OF THE CONTEXTUAL AND SEMANTIC RELEVANCE

Sentence	Contextual Relevance KG	Semantic Relevance (Attention Mechanism)
"User A remarked on Post Y: 'That's just great!'"	The KG links User A to their previous posts and connections, giving context that they usually use sarcasm.	The attention mechanism recognizes "great" and the emoji as vital signs of sarcasm, inferring the sentence as negative.
"User B posted: 'We need to be kind to each other.'"	The KG recognizes that User B habitually shares positive content, strengthening the positive context of this post.	The attention mechanism emphasizes the words "kind" and "each other," highlighting the post's positive moral intent.
"User C said: 'This is how we should treat persons by avoiding them.'"	The KG distinguishes that User C's post history comprises sarcastic or ironic reports.	The attention mechanism highlights "treat persons" and "avoid" as contradictory concepts, taking the sentence as sarcastic.
"User D responded: 'What a wonderful day to spread lies!'"	The KG links this post with other content by User D that has negative inferences.	The attention mechanism highlights "wonderful day" and "spread lies," detecting the sarcastic tone and negative intent.

TABLE 5. FEATURES IMPROVING IDENTIFICATION OF CONTEXTUAL AND SEMANTIC RELEVANCE

Features	Functionality	Role in Contextual Relevance	Role in Semantic Relevance
KG	Assimilates entities and relations, such as users, posts, and content types.	Delivers context by relating linked entities, and assists in understanding the relations within the data.	Improves the meaning of words by connecting them with associated ideas and entities.
Attention Mechanism	Emphases the most relevant parts of the input sequence during processing.	Selectively attends to main contextual indications in a post, certifying that the most pertinent data is recorded.	Assists the model in interpreting the nuances of the text by weighing the rank of diverse words and phrases.
BiLSTM Network	Processes input sequences in both forward and backward directions.	Keeps the context of the entire sequence by seeing data from both past and future words.	Captures the semantic meaning by understanding the flow of the text in both directions.
Entity Embeddings	Embeds users, posts, and other entities based on their relations in the KG.	Embeds entities in a manner that imitates their context, considering their relations to other entities.	Delivers a strong semantic illustration by embedding the meaning of words along with their associations.

TABLE 6. USER-GENERATED POSTS VERSUS USER-INTERACTED POSTS

Aspect	User-Generated Posts	User-Interacted Posts
Definition	Posts authored by users themselves.	Posts that users have interacted with (e.g., liked, remarked).
Main Focus	Direct expression of user feelings, attitudes, and intents.	Feedback and engagement with current content.
Contextual Analysis	Observe the content for inherent meanings and context.	Reflects how user interactions might modify the clarification of the post.
Semantic Analysis	Evaluates the content to detect potential immoral material based on its qualities.	Examines how interactions might change the perceived meaning or context of the post.
Role in Study	Central to detecting immoral content.	The secondary emphasis is to comprehend how engagement affects content insight.
Instance	A tweet stating a controversial view.	A tweet that received several likes and remarks.
Impact on Analysis	Directly marks the understanding of potentially harmful content.	Aids measure how user interactions might affect content interpretation.

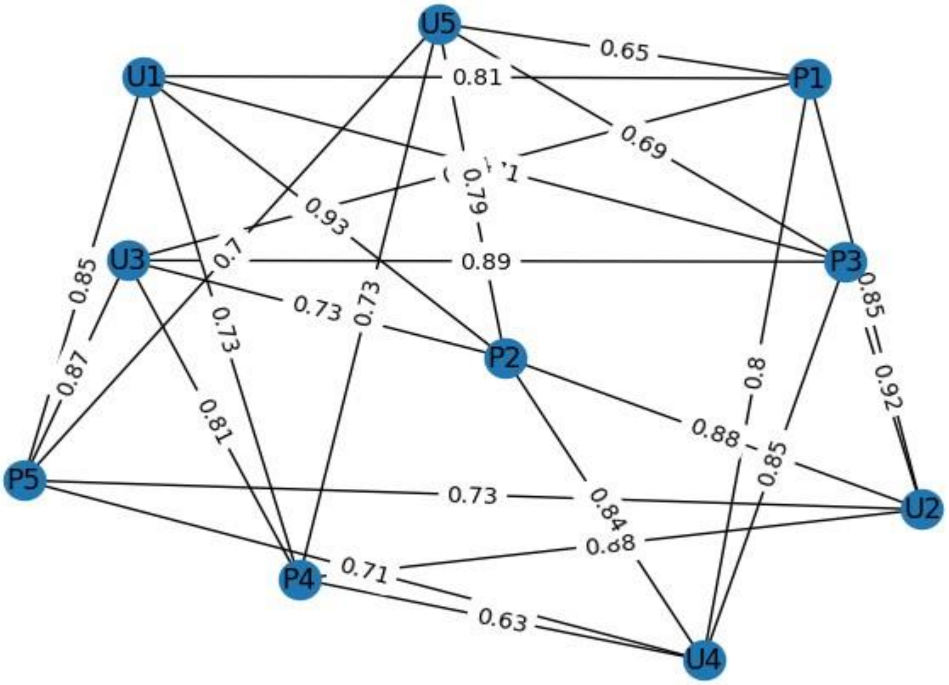

FIGURE 2. KG DESCRIBE THE RELATIONSHIP BETWEEN FIVE USERS AND THEIR POSTS

the one-hot encoded vectors with the embedding matrix is used to create word embeddings as shown in Eq.6.

$$Embed(w) = OneHot(w) \cdot E \tag{6}$$

The word embedding vector for the word w is Embed $(w) \in$ R^d. The Word2Vec is used in this study to increase the probability of correctly predicting context words from a target word.

$$Loss = -\sum_{context} log\, P(\ c \mid t\ ) \tag{7}$$

Where c indicates context words as shown in Eq.7 and target words are represented by t. The probability of context words provided via the target word is denoted by P (c|t).

2) GLOVE WORD EMBEDDING

The GloVe Word Embedding model attempts to train word vectors that encode semantic links between words based on co-occurrence statistics in a corpus. A logarithmic bilinear model is used in GloVe embedding. The logarithm operator is used to transform the ratio of co-occurrence probability values into differences [51]. Meaning therefore can be expressed as vector differences in log word vector space. In this study, the training method of the GloVe model focuses on a loss function called J, which is outlined in Eq.8.

$$J = \sum_{i,j=1}^{|V|} f(Y_{ij})\left(v_i^T \,.\, v_j + b_i + b_j - \log(X_{ij})\right)^2 \qquad (8)$$

$\sum_{i,j=1}^{|V|} i$ denotes the summarization of all word pairs i and j in the vocabulary. $f(Y_{ij})$ indicates a weighting function that estimates the relevance of the word pair (i, j) based on the number of times they appear together. $v_i^T \,.\, v_j$ represents the intersection of the word vectors v_i and v_j. $b_i and\ b_j$ is the Bias terminology related to the letters i and j. $\log(X_{ij})$ indicates the log of the observed word co-occurrence count between i and j. In the above equation, J is the overall goal function that must be reduced.

C. BILSTM NETWORK

BiLSTM is a form of Recurrent Neural Network (RNN) architecture used in NLP and sequence modeling. It enables the network to save knowledge from both the previous and the upcoming [52, 53]. To process the sequential data in both forward and backward directions concurrently, we employed the BILSTM network. By integrating the KG's retrieved features with the context learned by the BiLSTM networks. Through this bidirectional processing, the network can take into account both the prior and the upcoming at each time step, providing a more thorough knowledge of the sequence. The result of the BiLSTM at a specific time step t is obtained by joining the hidden forward and backward states.

$$h_t = [h_t^f, h_t^b] \qquad (9)$$

Eq.9 shows the joint hidden state at time step t is represented by h_t. The hidden state $h_t{}^f$ is at time step t from the forward LSTM. The hidden state represents $h_t{}^b$ at time step t from the reverse LSTM.

In this study, H indicates the hidden size of the BiLSTM, *X* represents the input sequence of word embeddings with *T* time steps, and *Y* is the output sequence. Eq.10 describes the forward and backward hidden states at each time step are concatenated and then passed through an output layer to produce the final output Y.

$$Y_t = softmax(W_y[\, h_t \,, \bar{h}_t] \; + \; b_y) \qquad (10)$$

W_y is the output weight matrix, b_y is the bias vector, and is [h_t, $\bar{h}_t$] represents the product of the forward and backward hidden states. These equations outline the BiLSTM network and show how output sequences are generated after input sequences are processed in both directions.

D. INTEGRATION OF KG FEATURES

Combining information produced from the graph with data from other sources is necessary to include KG features in a model. Let F_{Text}, F_{KG}, and F_{Other}, represents the textual, KG, and additional features. Eq.11 shows the nodes and edges of the graph act as the source of the KG's features.

$$F_{\mathrm{KG}} = \{ \mathrm{N}(v) \mid v \in V \} \cup \{ \mathrm{E}(e) \mid e \in E \} \qquad (11)$$

Where the embedding representations of nodes and edges are represented by N(v) and E(e). Word embedding is used to extract textual information from the input text as shown in Eq.12.

$$F_{\mathrm{Text}} = \{ \mathrm{E}(w) \mid w \} \qquad (12)$$

Where E (w) is the input text's word embedding vector for the word w. The task may also need the inclusion of other relevant features as defined by Eq.13.

$$F_{Other} = \{ F_1, F_2, \ldots \} \qquad (13)$$

Where F_i stands for a variety of extra features. Using a fusion technique, KG features are combined with additional features.

$$F_{\mathrm{Integrated}} = \mathrm{F}(\, \mathrm{F}_{\mathrm{KG}}, \mathrm{F}_{Text}, \mathrm{F}_{Other}) \qquad (14)$$

Where F is a fusion function that mixes several sets of features. The classification layer is provided with integrated characteristics after they have been gathered to make predictions. Here is a simplified illustration of how the classification layer might process integrated features. Let $F_{Integrated}$, represents an integrated feature achieved via fusing information from the KG, text, and other features as mentioned in Eq.14.

Classification Layer: The integrated features are input into the classification layer, which generates predictions as Eq.15 represents.

$$P = softmax(W_{class} F_{Integrated} + b_{class}) \qquad (15)$$

Where, P is the predicted class probabilities and W_{class} is the classification layer's weight matrix, bias vector, and b_{class}.

Loss Function: A suitable loss function that evaluates the difference between predicted probabilities and ground truth labels is used to train the model. To effectively reduce the loss function, we employed an Adaptive Moment Estimation (Adam) optimizer. It adjusts the learning rates based on the gradient history, enabling the model to converge more quickly and produce better results as Eq.16 represents.

$$Loss\ = loss(\ P,\ TL) \qquad (16)$$

Where TL refers to the labels that are attached to the input data.

Optimization: To reduce the loss function, the optimization method modifies the model's parameters as presented in Eq.17.

$$Update\ Rule : \theta \leftarrow \theta - \alpha \frac{\partial Loss}{\partial \theta} \qquad (17)$$

Where θ represents the parameters of the model, $\frac{\partial \text{Loss}}{\partial \theta}$ denotes the gradient of the loss regarding parameters, and α represents the learning rate.

In the proposed work, contextual and semantic relevance perform an important role in correctly identifying immoral content. Table 4 delivers example sentences that explain the contextual and semantic relevance of posts. Table 5 represents features improving the identification of contextual and semantic relevance. Therefore, a more detailed discussion on how each feature explicitly contributes to and improves these aspects, with practical instances from the datasets, would deliver a simplified understanding of their effect. For example, clarifying how the KG disambiguates the meaning of a sarcastic post or how the attention mechanism chooses the most contextually significant words could add depth to the study's analysis. The proposed study highlights the contextual and semantic analysis of posts, which requires distinguishing between user-generated posts and user-interacted posts. This difference permits us to precisely evaluate how each kind of post adds to the understanding of immoral content and its implications. Table 6 represents the relevant concept associated with user-generated posts versus user-interacted posts, concentrating on their contextual and semantic clarification.

IV. EXPERIMENTS

In this section, the overall experimental findings are reported and discussed. We also describe the training process, experimental setting, and evaluation metrics to show the usefulness of the proposed framework.

A. TRAINING PROCESS

We cautiously chose the benchmark datasets SARC and HatEval throughout the training period to make sure they were appropriate for with proposed technique. To enhance the quality of the data and get it ready for well-organized model training, this necessary widespread data pretreatment measures. We tokenized the text into words, decoded it to lowercase for reliability, and then accomplished a systematic data cleaning process to eliminate noise such as URLs and special characters. We used methods containing eliminating frequently occurring stop words, lemmatizing words for word normalization, and using oversampling or undersampling techniques to decide data imbalance. Besides, we employed label encoding to simplify categorical labels into numerical values and truncated sequences as required to standardize the length of the orders. After the data preprocessing, we ongoing by taking an appropriate architecture before training our model. We put into practice our proposed BiLSTM network, which depends on KG and integrates KG features for improved immoral post detection on social media. We interpreted the textual input into numerical illustrations using pre-trained word embeddings such as GloVe, capturing semantic meanings necessary for accurate classification. We derive beneficial feature representations, like node embeddings, edge embeddings, and graph embeddings, by utilizing the built KG, which shows associations between users, posts, and content kinds. To enhance the contextual data from user-generated texts, we trained the BiLSTM network to process sequential input in both forward and backward orders instantaneously. To develop detailed illustrations for individual text, we joined KG features with textual features and any other features using fusion methods. In this study, we split the dataset into 80% for training and 20% for testing while maintaining the same class distribution across both sets. This approach reflects the natural imbalance of the dataset's classes, such as moral vs. immoral and sarcastic vs. non-sarcastic posts. By preserving the class proportions, we ensure that the model learns from a representative distribution and can be accurately evaluated on how well it handles real-world imbalances in content moderation tasks. To maximize model performance, we adjusted hyperparameters containing learning rate, BiLSTM hidden size batch size, and number of epochs throughout the training phase. Employing the Adam optimizer together with a suitable loss function such as categorical cross-entropy, we trained the model to decrease the inconsistency between the ground truth labels and projected possibilities. To make sure the model is efficient in observing immoral posts, we thoroughly evaluated its performance after training using assessment metrics like accuracy, precision, recall, and F1-score on a validation set. After the validation outcomes are satisfactory, we finally put the model into production for usage in practice. We strictly observe the proposed model performance and regularly reskill it to adjust to changing fashions and data allocations.

B. MODEL PARAMETERS

This section describes the simulation tools used in the proposed study. An NVIDIA GeForce GTX GPU is used to expedite processing operations, and an Intel Core i7 CPU with 16 GB of RAM is part of the computational setup. The research makes use of two benchmark datasets, SARC and HatEval, which provide annotations for posts that are moral or immoral, sarcastic or not, with a total of 12,000 and 57,000 instances, respectively. The main programming language is Python 3.6, which is combined with TensorFlow for deep learning applications. Pandas and NumPy are used for data manipulation and analysis, and NetworkX is used for graph processing. Text preprocessing is based on NLTK.

TABLE 7. MODEL PARAMETERS AND EXPERIMENTAL SETUP

Parameter	Specifications
CPU	Intel Core i7
RAM	16 GB
GPU	NVIDIA GeForce GTX
Dataset	Benchmark Dataset (SARC and HatEval)
Label	Moral and Immoral
Instances	12000 and 57000
Programming Language	Python (version 3.6)
Deep Learning Framework	TensorFlow
Data Manipulation and Analysis	Pandas, NumPy
Graph Processing	NetworkX
Text Preprocessing	NLTK
Dimensionality Reduction	Scikit-learn
Evaluation Metrics	Accuracy, Precision, Recall, F1-Score
Word Embeddings	Word2Vec, GloVe
Model Input	Word Embeddings
Model Output	Moral= 0, Immoral= 1
Learning_rate	0.001-0.01
Model Optimizer	Adam
Batch size, number of epochs	16-64
ReLU	Activation Function

Scikit-learn helps in the reduction of dimensionality. The evaluation metrics accuracy, precision, recall, and F1-score are employed to evaluate the performance of the proposed model. Table 7 represents the experimental setup and model parameter summary. Word2Vec and GloVe embedding improves the model's comprehension of textual material. Word embedding is the model's input, and its output is a prediction of moral and immoral labels (0 and 1).

During model training, hyperparameters such as the number of epochs, batch size (16-64), learning rate (0.001-0.01), and the Adam optimizer are optimized. To add non-linearity and enhance model performance, the neural network architecture makes use of the Rectified Linear Unit (ReLU) activation function. The SARC and HatEval datasets are primarily focused on English language content. We strive to ensure that the training data is as balanced and inclusive as possible. To mitigate biases, the proposed mode uses re-sampling, re-weighting, and adversarial debiasing in the training data. The SARC dataset focused on sarcastic comments. It presents challenges because sarcasm is often subtle and context-dependent. It makes it difficult for the model to effectively recognize the immoral content. We need to update and fine-tune the model again and again. The HatEval dataset, which deals with hate speech, is more straightforward as the content typically conveys clear intent. However, this clarity can limit the model's ability to generalize to more nuanced or ambiguous cases, such as those found in the SARC dataset.

C. EVALUATION METRICS

This section demonstrates the accomplishment matrices and evaluation standards that are applied throughout the study process to ensure the reliability of the findings. The accuracy, precision, recall, and F1-score are employed to evaluate the efficiency of the proposed model. The accuracy represents the overall percentage of cases in the dataset that are properly classified. It provides an evaluation of the model's accuracy as Eq.18 represents.

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \tag{18}$$

To determine how many of the predicted immoral posts are truly immoral, Precision analyzes the ratio of true positives to all predicted positives. To determine how many of the predicted immoral posts are truly immoral, Precision analyzes the ratio of true positives to all predicted positives as shown in Eq.19.

$$P = \frac{TP}{TP + FP} \tag{19}$$

Recall, also referred to as sensitivity or true positive rate, and quantifies the percentage of real immoral posts that the model accurately identified as Eq. 20 represents.

$$R = \frac{TP}{TP+FN} \tag{20}$$

The F1-score is a balanced indicator of the model's accuracy because it is the harmonic mean of precision and recall as Eq.21 signifies.

$$F_{score} = \frac{2 \times P \times R}{P + R} \tag{21}$$

Where $\mathrm{T_P}$ shows true positive values, while $\mathrm{T_N}$ signifies true negative values. Moreover, $\mathrm{F_P}$ and $\mathrm{F_N}$ denotes false positive and false negative values respectively. While P stands for precision and R for recall.

V. RESULTS AND DISCUSSIONS

This section provides the results of the proposed study using the KG-based BiLSTM technique. Table 8 indicates the overall results obtained using BILSTM and KG-based BILSTM.

TABLE 8. RESULTS OBTAINED USING BiLSTM AND KG-BASED BILST

Model	Dataset	Precision (%)	Recall (%)	F1-Score
BiLSTM	SARC	76.12	78.45	77.27
BiLSTM	HatEval	73.62	78.24	75.86
KG-based BiLSTM	SARC	82.36	83.23	82.79
KG-based BiLSTM	HatEval	78.75	90.13	84.06

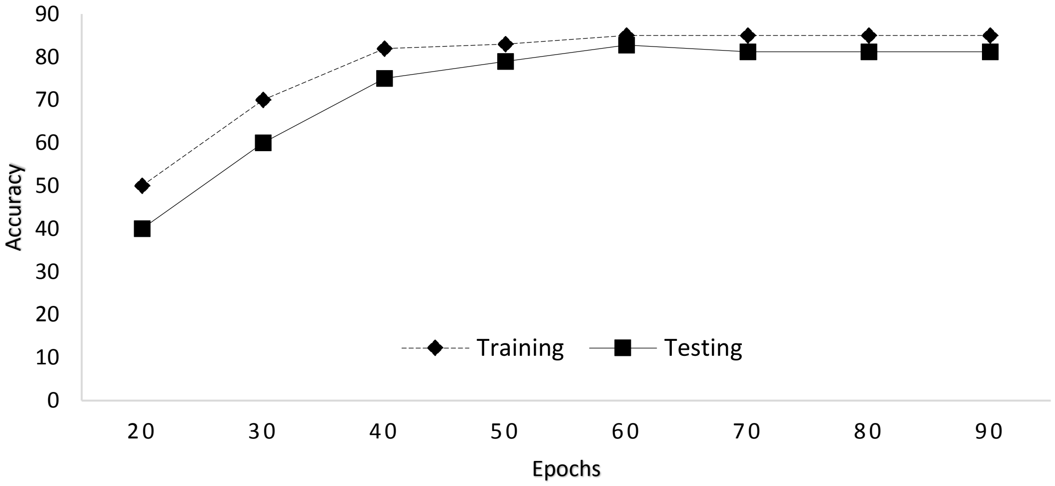

FIGURE 3. TRAINING AND TESTING ACCURACY ON THE SARC DATASET

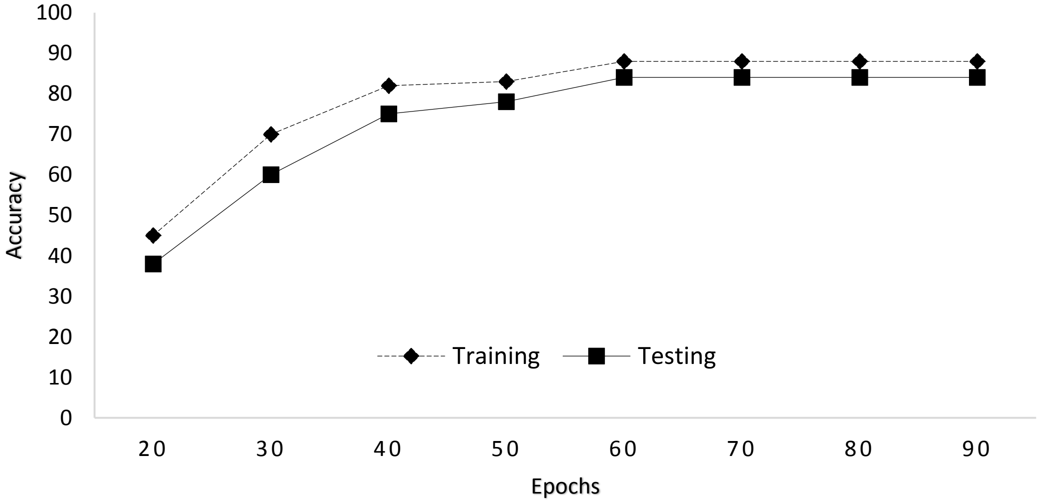

FIGURE 4. TRAINING AND TESTING ACCURACY ON THE HATEVAL DATAS

TABLE 9. COMPARISON RESULTS USING DIFFERENT MODELS FOR SARC, HATEVAL DATASETS

Study	Method	Dataset	Precision (%)	Recall (%)	F1- Score (%)
[40]	MHA-BiLSTM	SARC Dataset	60.26	53.71	56.79
[10]	RSGNN	HatEval Dataset	74.29	74.14	74.04
Proposed	KG-Base BiLSTM	SARC	82.36	83.23	82.79
		HatEval	78.75	90.13	84.06

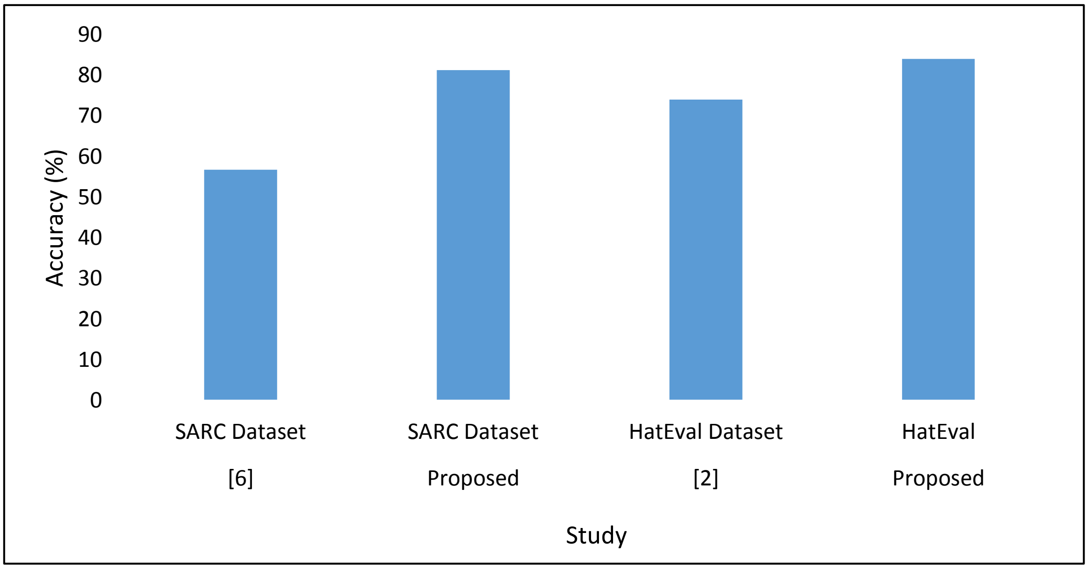

FIGURE 5. COMPARISON WITH CLOSELY RELATED WORKS

This study highlights the operation of the proposed method in noticing immoral posts on social media. It represents important developments in text classification perfection and strength compared to traditional approaches. KG-based BiLSTM provides the highest precision, recall, and F1-score than BiLSTM on the SARC and HatEval datasets. The effectiveness of the model in properly identifying immoral and non-immoral postings based on its predictions is measured using these evaluation criteria. Table 9 represents the comparison of the proposed model with the state-of-the-art techniques. The author of this study employed the Attention-Based Bidirectional Long-Short Memory (MHA-BiLSTM) network to find sarcastic remarks in a given corpus. The model utilized in this study had an F1 score of 56.79% on the SARC dataset [40]. This study's author created a Relation- Special Graph Neural Network (RSGNN) for abusive post-detection. The experiment demonstrates that their model achieved an F1-score of 74.04% on the HatEval dataset [10]. The proposed KG-Base BiLSTM on the SARC dataset achieved a remarkable precision of 82.36% and a recall of 83.23%. The proposed KG-Base BiLSTM on the HatEval dataset also performed best with precision, and recall of 78.75%, and 90.13% respectively. To summarize the discussion, we compare the performance of various studies on the SARC and HatEval datasets, the proposed KG-Based BiLSTM network outperformed the other models and achieved the highest F1-score of 82.79% and 84.06% respectively. Figure 3 shows training and testing accuracy on the SARC dataset. Figure 4 indicates training and testing accuracy on the HatEval dataset. Both graphs represent that the proposed model achieved more efficient results than the baseline model on the same dataset. We integrated a KG to capture the associations between diverse entities and ideas. We compare the results of the proposed methodology with those of the best methods using the SARC and HatEval datasets. Figure 5 represents the comparison of the proposed model with closely related works. This assists in understanding the context and semantics of the posts more precisely. By embedding the associations between users, posts, and content types, the KG improves the model's capability to identify immoral content with superior results. The proposed study influences an Attention-Based BiLSTM network. The attention mechanism permits the model to focus on the most applicable parts of the text, refining its capability to capture essential contextual data. The bidirectional nature of the BiLSTM allows the model to reflect context from both directions in the text, improving its understanding of nuances and enhancing identification performance. We have adapted the feature extraction procedure to better control contextual meanings and language nuances. This comprises purifying how we process and understand informal terms, which are shared in social media posts. We employed definite training approaches and fine-tuning methods to optimize the proposed model's performance. This involves adjusting hyperparameters and using methods like dropout and regularization to avoid overfitting and certify well generalization. The model generally performed better on the HatEval dataset because hate speech is often more explicit and easier to identify, with clear markers of offensive language.

In contrast, the SARC dataset, which focuses on sarcasm, presented more challenges due to the subtle and context-dependent nature of sarcastic comments, making it harder for the model to accurately detect them. These differences highlight the varying complexities of detecting different types of immoral content, with sarcasm requiring a more nuanced understanding compared to hate speech.

Currently, the proposed model is focused on detecting immoral posts in English, as the SARC and HatEval datasets are English-based. Its performance in other languages has not been evaluated yet. However, in future work, we plan to extend the model's capabilities to detect immoral posts in other languages, such as Roman Urdu and Pashto, which will involve adapting and retraining the model to handle the unique linguistic and cultural nuances of these languages. For the detailed evaluation of the proposed model, we also calculate the Cohen's Kappa metric. It measures the agreement between the actual labels, adjusted for chance agreement and the model's calculations. The Cohen's Kappa score obtained for the SARC dataset is 0.72. Using the HatEval dataset we obtained the Cohen's Kappa score of 0.70.

The adoption of KG assists in quickly contextualizing new information, enhancing the model's ability to detect immoral posts in real-time. The use of attention mechanisms enables the model to efficiently focus on the most relevant aspects of content improving the detection process. The primary factors inducing time complexity are the size of the datasets (12000 and 57000 instances for SARC and HatEval, correspondingly), the number of epochs (16-64), and the complexity of the BiLSTM and attention mechanisms. The BiLSTM models regularly have a time complexity of O (n*d*h), whereas *n* is the sequence length, *d* is the embedding dimension, and *h* is the hidden units, the proposed model is computationally demanding, particularly during the training period. The attention mechanism, which includes computing position scores, further enhances the time complexity. The usage of GloVe and Word2Vec embeddings, which have their computational costs during the embedding lookup and processing, also contributes to the overall time complexity. The space complexity is mainly determined using the model's parameters and the size of the datasets. Table 10 shows the time and space complexity of the proposed method. The BiLSTM model needs important memory for storing weights and intermediate states, particularly with 16 GB of RAM. The KG processing, which is handled by the NetworkX library, adds to the space complexity due to the need to store and operate graph data structures. The embeddings (GloVe and Word2Vec) also contribute to the space requirements, as they need memory to store pre-trained vectors. The proposed technique's space complexity is consequently dependent on the size of the input data, the dimensionality of the embeddings, and the number of parameters in the BiLSTM and attention layers.

TABLE 10. TIME AND SPACE COMPLEXITY

Component	Time Complexity	Space Complexity
BiLSTM Model	O(n * d * h)	O(h * d)
Attention Mechanism	O(n^2 * h)	O(n * h)
KG Processing (NetworkX)	O(k) to O(k^2)	O(k . d)
GloVe/Word2Vec Embedding Lookups	O(n * d)	O(V * d)
Data Preprocessing (NLTK, Scikit-learn)	O(n * log(n))	O(n * d)
Model Training (Overall)	O(n * epochs * model complexity)	O(parameters + data size)

VI. CONCLUSIONS AND FUTURE DIRECTIONS

This work provides a novel method for identifying immoral posts on social media platforms, using KG and Attention-Based BiLSTM works. The content of the SN post is parsed to identify entities and relationships present in the text. In the context of immoral post detection, the KG assists us in recognizing the content of posts as harmful, unethical, or violated by representing it in a structured format. It provides the relationship among various entities and keywords or phrases associated with immoral content. If the post content is highly similar to immoral content, it is marked for further analysis. Incorporating Attention-based BiLSTM improved detection performance, mainly in controlling the complex nuances of sarcasm and moral findings in the text. It processes the sequences in both forward and backward directions, allowing the model to capture context from both past and future tokens in a sequence. The attention mechanism enables the model to focus on the input sequences that are indicative of immoral behavior. The proposed model is validated using publically available benchmark datasets such as SARC and HatEval. This work focuses on commonly recognized awkward behaviors such as hate speech and sarcasm, which are utilized to hurt or demean others. This work leverages a data-driven approach exploiting KG and attention-based BiLSTM, which can be adjusted to various contexts by integrating culturally explicit knowledge into the model. There is still an opportunity for future research into customized moderation systems, in which users may define their content filters depending on their preferences. Future research may include this individualized dimension through user feedback, which will change the system's sensitivity to specific sorts of content. Furthermore, in the future, we want to incorporate Graph Neural Networks (GNNs) to model the relationships between different entities within a post, such as users, topics, and sentiments.

Acknowledgments: This study is supported via funding from Prince Sattam bin Abdulaziz University project number (PSAU/2024/R/1445).

Authors Contributions: Khan gives the idea of the proposed work and thoroughly investigates the related work section. Atta done overall inspection and contribution, including related work, proposed work, and experimentation. Sajid provides technical setup and experimentation. Mohammed thoroughly investigated the paper and gave suggestions for enhancement. Wahab reviewed the complete paper, gave suggestions for introduction, and proposed work. Ashraf provides suggestions for experimentation. Finally, all authors read, and review the paper.

Availability of Data and Materials: The datasets used in this paper were obtained from Kaggle. Available online via the following links: https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset, and https://www.kaggle.com/datasets/feyzazkefe/olid-dataset.

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

REFERENCES

1. Ashraf, N., A. Zubiaga, and A. Gelbukh, *Abusive language detection in youtube comments leveraging replies as conversational context.* PeerJ Computer Science, 2021. **7**: p. e742.
2. Lee, H.-S., et al., *An abusive text detection system based on enhanced abusive and non-abusive word lists.* Decision Support Systems, 2018. **113**: p. 22-31.
3. Putri, N.L.P.M.S., D. Nurjanah, and H. Nurrahmi, *Cyberbullying detection on twitter using support vector machine classification method.* Building of Informatics, Technology and Science (BITS), 2022. **3**(4): p. 661− 666-661− 666.
4. Aljero, M.K.A. and N. Dimililer, *A novel stacked ensemble for hate speech recognition.* Applied Sciences, 2021. **11**(24): p. 11684.
5. Shah, F., et al., *Artificial Intelligence as a Service for Immoral Content Detection and Eradication.* Scientific Programming, 2022. **2022**.
6. Chen, Y., et al. *Detecting offensive language in social media to protect adolescent online safety*. in *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*. 2012. IEEE.
7. Van Hee, C., et al. *Detection and fine-grained classification of cyberbullying events*. in *Proceedings of the international conference recent advances in natural language processing*. 2015.
8. Wulczyn, E., N. Thain, and L. Dixon. *Ex machina: Personal attacks seen at scale*. in *Proceedings of the 26th international conference on world wide web*. 2017.
9. Kapil, P., A. Ekbal, and D. Das, *Investigating deep learning approaches for hate speech detection in social media.* arXiv preprint arXiv:2005.14690, 2020.
10. Song, R., et al., *Improving Abusive Language Detection with online interaction network.* Information Processing & Management, 2022. **59**(5): p. 103009.
11. Govindan, V. and V. Balakrishnan, *A machine learning approach in analysing the effect of hyperboles using negative sentiment tweets for sarcasm detection.* Journal of King Saud University-Computer and Information Sciences, 2022. **34**(8): p. 5110-5120.
12. Bhardwaj, A., *Sentiment Analysis and Text Classification for Social Media Contents Using Machine Learning Techniques.* Available at SSRN 3735851, 2020.
13. Anand, M., et al., *Deep learning and natural language processing in computation for offensive language detection in online social networks by feature selection and ensemble classification techniques.* Theoretical Computer Science, 2023. **943**: p. 203-218.
14. Kshirsagar, R., et al., *Predictive embeddings for hate speech detection on twitter.* arXiv preprint arXiv:1809.10644, 2018.
15. Sundaram, A., et al., *A Systematic Literature Review on Social Media Slang Analytics in Contemporary Discourse.* IEEE Access, 2023. **11**: p. 132457-132471.
16. van Huijstee, M., et al., *Harmful Behaviour Online: An investigation of harmful and immoral behaviour online in the Netherlands.* 2022.
17. Franco, M., O. Gaggi, and C.E. Palazzi. *Analyzing the use of large language models for content moderation with chatgpt examples*. in *Proceedings of the 3rd International Workshop on Open Challenges in Online Social Networks*. 2023,p.1-8, https://doi.org/10.1145/3599696.3612895.
18. Jhaver, S., et al., *Personalizing content moderation on social media: User perspectives on moderation choices, interface design, and labor.* Proceedings of the ACM on Human-Computer Interaction, 2023. **7**(CSCW2): p. 1-33. doi: 10.1145/3610080.
19. Farías, D.I.H., et al. *Valento: Sentiment analysis of figurative language tweets with irony and sarcasm*. in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. 2015.
20. Rajamanickam, S., et al., *Joint modelling of emotion and abusive language detection.* arXiv preprint arXiv:2005.14028, 2020.
21. Saumya, S., A. Kumar, and J.P. Singh, *Filtering offensive language from multilingual social media contents: A deep learning approach.* Engineering Applications of Artificial Intelligence, 2024. **133**: p. 108159.
22. Mazari, A.C. and H. Kheddar, *Deep learning-based analysis of Algerian dialect dataset targeted hate speech, offensive language and cyberbullying.* International Journal of Computing and Digital Systems, 2023.
23. Miao, Z., et al., *Detecting Offensive Language on Social Networks: An End-to-end Detection Method based on Graph Attention Networks.* arXiv preprint arXiv:2203.02123, 2022.
24. Thaokar, C., et al., *N-Gram based sarcasm detection for news and social media text using hybrid deep learning models.* SN Computer Science, 2024. **5**(1): p. 163.
25. Fan, X., et al., *Identifying Hate Speech Through Syntax Dependency Graph Convolution and Sentiment Knowledge Transfer.* IEEE Access, 2023.
26. Vasantharajan, C. and U. Thayasivam, *Towards offensive language identification for Tamil code-mixed YouTube comments and posts.* SN Computer Science, 2022. **3**(1): p. 94.
27. Yasaswini, K., et al. *IIITT@ DravidianLangTech-EACL2021: Transfer Learning for Offensive Language Detection in Dravidian Languages*. in *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. 2021.
28. Nobata, C., et al. *Abusive language detection in online user content*. in *Proceedings of the 25th international conference on world wide web*. 2016.
29. Ibrohim, M.O. and I. Budi, *A dataset and preliminaries study for abusive language detection in Indonesian social media.* Procedia Computer Science, 2018. **135**: p. 222-229.

30. Iqbal, F., et al., *A hybrid framework for sentiment analysis using genetic algorithm based feature reduction.* IEEE Access, 2019. **7**: p. 14637-14652.

31. Davidson, T., D. Bhattacharya, and I. Weber, *Racial bias in hate speech and abusive language detection datasets.* arXiv preprint arXiv:1905.12516, 2019.

32. Schmidt, A. and M. Wiegand. *A survey on hate speech detection using natural language processing*. in *Proceedings of the fifth international workshop on natural language processing for social media*. 2017.

33. Sapountzi, A. and K.E. Psannis, *Social networking data analysis tools & challenges.* Future Generation Computer Systems, 2018. **86**: p. 893-913.

34. Abokhodair, N. and S. Vieweg. *Privacy & social media in the context of the Arab Gulf*. In Proceedings of the 2016 ACM Conference on Designing Interactive Systems (DIS '16). Association for Computing Machinery, Brisbane, QLD, Australia, p.672–683. doi: 10.1145/2901790.2901873.

35. Franco, M., et al. *A technology exploration towards trustable and safe use of social media for vulnerable women based on islam and arab culture*. in *Proceedings of the 2022 ACM Conference on Information Technology for Social Good*. 2022. 138–145. doi: 10.1145/3524458.3547259.

36. d'Sa, A.G., I. Illina, and D. Fohr. *Bert and fasttext embeddings for automatic detection of toxic speech*. in *2020 International Multi-Conference on:"Organization of Knowledge and Advanced Technologies"(OCTA)*. 2020. IEEE.

37. Sharma, D.K., et al., *Sarcasm detection over social media platforms using hybrid auto-encoder-based model.* Electronics, 2022. **11**(18): p. 2844.

38. Mossie, Z. and J.-H. Wang, *Social network hate speech detection for Amharic language.* Computer Science & Information Technology, 2018: p. 41-55.

39. Sap, M., et al. *The risk of racial bias in hate speech detection*. in *ACL*. 2019.

40. Kumar, A., et al., *Sarcasm detection using multi-head attention based bidirectional LSTM.* Ieee Access, 2020. **8**: p. 6388-6397.

41. Struß, J.M., et al., *Overview of germeval task 2, 2019 shared task on the identification of offensive language.* 2019.

42. Romim, N., et al., *HS-BAN: A Benchmark Dataset of Social Media Comments for Hate Speech Detection in Bangla.* arXiv preprint arXiv:2112.01902, 2021.

43. Kumar, R., et al. *Benchmarking aggression identification in social media*. in *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*. 2018.

44. Raza, M.O., et al. *Detecting cyberbullying in social commentary using supervised machine learning*. in *Advances in Information and Communication: Proceedings of the 2020 Future of Information and Communication Conference (FICC), Volume 2*. 2020. Springer.

45. Subramanian, M., et al., *Offensive language detection in Tamil YouTube comments by adapters and cross-domain knowledge transfer.* Computer Speech & Language, 2022. **76**: p. 101404.

46. Serra, J., et al. *Class-based prediction errors to detect hate speech with out-of-vocabulary words*. in *Abusive Language Workshop*. 2017. Abusive Language Workshop.

47. Fang, Y., et al., *Dynamic knowledge graph based fake-review detection.* Applied Intelligence, 2020. **50**: p. 4281-4295.

48. Lin, J., *Leveraging World Knowledge in Implicit Hate Speech Detection.* arXiv preprint arXiv:2212.14100, 2022.

49. Agrawal, G., et al., *Building Knowledge Graphs from Unstructured Texts: Applications and Impact Analyses in Cybersecurity Education.* Information, 2022. **13**(11): p. 526.

50. Wang, S., W. Zhou, and C. Jiang, *A survey of word embeddings based on deep learning.* Computing, 2020. **102**: p. 717-740.

51. Soto, C.P., et al., *Application-specific word embeddings for hate and offensive language detection.* Multimedia Tools and Applications, 2022. **81**(19): p. 27111-27136.

52. Yin, J., et al., *Forecast of short-term daily reference evapotranspiration under limited meteorological variables using a hybrid bi-directional long short-term memory model (Bi-LSTM).* Agricultural Water Management, 2020. **242**: p. 106386.

53. Kamyab, M., G. Liu, and M. Adjeisah, *Attention-based CNN and Bi-LSTM model based on TF-IDF and glove word embedding for sentiment analysis.* Applied Sciences, 2021. **11**(23): p. 11255.

BIBI SAQIA received the MS degree in Computer Science from University of Science and Technology Bannu in 2018. Currently her PhD is in progress from University of Science and Technology Bannu. She has more than 7 publications in various reputed journals and conferences including IEEE Transactions. Her research interest includes Data Mining, Machine Learning and Artificial Intelligence.

KHAIRULLAH KHAN received the Ph.D. degree in information technology from UniversitiTeknologi PETRONAS, Malaysia, in 2012, where he worked on machine learning for the automatic detection of opinion targets from text. He is currently an Professor with the Department of Computer Science, University of Science and Technology, Bannu, Pakistan. He has more than 45 publications in various reputed journals and conferences including IEEE Transactions. His research interest includes Data Mining, Web Mining, Opinion Mining, Machine Learning and Artificial Intelligence.

ATTA UR RAHMAN received the MS degree in Computer Science from University of Science and Technology Bannu in 2018, and the Ph.D. degree in Computer Science from Ghulam Ishaq Khan Institute of Engineering Sciences and Technology in 2022. Currently, he is working as an Assistant Professor at Riphah Institute of System Engineering (RISE), Riphah International University Islamabad, Pakistan. He has more than 15 publications in various reputed journals and conferences including IEEE Transactions. His research interest includes Human-computer Interaction, Artificial Intelligence in healthcare, and Federated learning for privacy preserving.

WAHAB KHAN received the M.S. degree in computer science from the University of Science and Technology, Bannu, Pakistan, in 2009. He is currently pursuing the Ph.D. degree in computer science with International Islamic University Islamabad, Pakistan. He has more than 20 publications in various reputed journals and conferences including IEEE Transactions. His research interests include natural language processing, machine learning, and deep learning and data mining.

SAJID ULLAH KHAN earned his Ph.D. degree from the University Malaysia Sarawak, Malaysia in 2016. His research interests encompass image analysis and understanding using machine learning and deep learning models. With nearly 13 years of experience, he has worked in various government and public organizations, engaging in both teaching and research roles. Presently, he serves as an Associate Professor in the Information Systems Department at the College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, Alkharj, KSA. He has contributed to journal papers, books, and conference proceedings, and has actively participated in talks and seminars.

MOHAMMED ALKHOWAITER Assistant professor in Prince Sattam bin Abdulaziz University Alkharj KSA. His area of interest is Information security, information retrieval, NLP.

ASHRAF ULLAH received the MS degree in Computer Science from University of Science and Technology Bannu in 2010. Currently his PhD is in progress from University of Science and Technology Bannu. He has different publications in various reputed journals and conferences including IEEE Transactions. His research interests include natural language processing, machine learning, and deep learning and data mining.