DETECTION OF CYBERBULLYING IN SOCIAL NETWORKS: COMPARING MACHINE LEARNING AND TRANSFER LEARNING METHODS

[1] Mahadevuni Madhubabu, [2] Dr. K. Santhi Sree

[1] Student, Department of Information Technology, University College of Engineering, Science and Technology, JNTUH Hyderabad

[2] Professor, Department of Information Technology, University College of Engineering, Science and Technology, JNTUH Hyderabad

Abstract: Information and Communication Technologies have revolutionized communication, but cyberbullying poses severe challenges, demanding automated solutions for effective detection on social media platforms. The PROJECT emphasizes feature extraction and selection techniques to enhance the model's understanding of cyberbullying instances, incorporating both traditional Machine Learning and Transfer Learning approaches. The project leverages diverse models, including LinearSVC, Logistic Regression, DistilBert, DistilRoBerta, and Electra, encompassing Machine Learning, Transfer Learning, and Deep Learning methodologies for robust cyberbullying detection.

Index terms - *Cyberbullying detection, DistilBert, machine learning, pre-trained language models (PLMs), transfer learning, toxicity features, AMiCa dataset, LIWC, empath.*

1. INTRODUCTION

Information and Communication Technologies (ICT) have become an integral part of everyone's life, evolving imperceptibly with time, catalyzing online communication between people. Communication has been just one button click with the widespread use of the online platform, facilitating the growth of social networking. ICT dominance has a dark side when people easily misuse technological advancement with abusive behaviors such as cyberbullying. Cyberbullying is the expanded form of direct or traditional bullying through electronic platforms [1], [2], [3], [4], [5], [6]. Social media becomes the virtual medium for bullying, shielding the bully's identity, making detecting cyberbullying a complex and challenging mission to protect online communities.

Cyberbullying cases increase with volumized Internet usage because it can be easily committed anonymously [7], leading to a grave public health concern that brings many negative impacts, such as mental, psychological, and social problems [8]. While cyberbullying victims tend to suffer from mental health problems such as depression, anxiety, loneliness, and anhedonia, some are reported to be committing self-injurious behavior and suicidal ideation [9]. Initially, the community implemented a manual approach to monitoring cyberbullying activities. Parent-Teacher Association started a good initiative from the Japanese school that formed Internet Patrol to help filter websites manually with inappropriate content, but it is impossible to handle the vast volume of data on the Internet within a short time without a computational approach [10], [11], [12].

Automating cyberbullying detection is paramount to facilitate the process, ensuring a safe environment within online social media. As the computational text analysis can effectively be adopted to examine the social and cultural phenomena [13], the primary focus of this research is to automate the detection of cyberbullying instances from the unruly post, deeming the problem as a text classification task with the help of stateof-the-art techniques using artificial intelligence and natural language processing knowledge. By natural language processing, text classification is frequently employed in identifying the category of a given corpus through several stages, such as text preprocessing, feature extraction, and the development of a classification model [14].

Social media companies have developed policies and mechanisms to maintain the regulation of social media platforms. However, the social media company was not performing well in tackling cyberbullying [15], [16]. The available mechanisms are usually user-dependent, requiring users to report content, block, or unfriend, a passive way of mitigating cyberbullying [17]. Although the implementation of algorithms with supervised machine learning works to detect cyberbullying events and helps to expunge posts that may contain foul words; however, the outcome is not as accurate as those reported by users [17]. Furthermore, metadata associated with the online platform and user information are not always available due to privacy protection [18], [19]. In that case, textual content posted by the online platform users is the base input for cyberbullying detection model [20].

The initial studies on automatic cyberbullying detection deemed the presence of ‘‘bad’’ words (insult and swear words) or profane terms to be one factor in making a post likely to be an act of cyberbullying. However, looking for a list of words to detect such events is not very effective because the words or sentences can be easily deformed or obfuscated in terms of spelling, and a consistent list update is required [21]. Using textual features such as the presence of ‘‘bad’’ words (insult, swear, profane word) in making a post to be an act of cyberbullying has its limitation since the explicit existence of these words is not always right to detect cyberbullying [22]. Extraction of additional features by expanding the usual bagof-words text representation is needed to improve the performance of cyberbullying detection model [18].

2. LITERATURE SURVEY

Cyberbullying among Turkish high school students. Scandinavian Journal of Psychology. Cyberbullying, a new form of the traditional bullying that has been transferred to the electronic environments (social media, online gaming environments, blogs, etc.), from the physical context to the virtual context, refers mainly to aggression that is deliberately carried out by adolescents. This study [1] aims to measure the level of cyberbullying in Turkish high school students living in Eastern Turkey and identify the demographic and socioeconomic factors which lead to being bully and being cyberbullied. The study population consists of 470 students aged from 15-19 years. exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) were implemented to identify the factor structure of the scale and it was observed that the Turkish version of the cyberbullying scale (CBS) is best represented by a one-factor structure. The comparisons across demographic and socioeconomic variables were implemented using independent samples t test, one-way ANOVA, and Tukey HSD. To summarize the key findings, the variables that significantly affect the students' CBS scores are; gender, school type, number of siblings, ownership of a mobile phone, length of ownership of a mobile phone, private access to the Internet, family supervision, purpose of Internet usage, length of time spent on the Internet and type of application used to message with others [34,35].

This study is conducted to learn about experiences and practices to cope with cyberbullying among high school students in Hanoi and to explore the association between the average time of Internet used per day among high school students in Hanoi, Vietnam, and the risk of being cyberbullied. A total of 215 students aged 13–18 years completed an online survey using respondent-driven sampling method [2]. The experience of being cyberbullied was examined using the modified Patchin and Hinduja's scale. The prevalence of experiencing at least one type of cyberbullying was 45.1%. The most common type of cyberbullying was being called by names/made fun of. The average daily time spent on Internet showed dose-response association with the risk of being cyberbullied. The prevalence of having experienced cyberbullying was 54% among subjects who used Internet >3 hours/day compared to 39% among those who used 1–3 hours and 30% among those who used <1 hour. In terms of practices to cope with this, most students chose to ignore it and not share information with their family or teacher. The most frequent method to overcome this problem was talking with friends (60.8%). Research shows that the prevalence of cyberbullying victimization in Hanoi was high, and student's practices to cope with this new form of bullying were not efficient. Online time had dose-response association with risk of cyberbullying [34,35]. More attention is needed to increase level of society/school awareness to prevent cyberbullying in Hanoi.

Bullying is the deliberate physical and psychological abuse that a child receives from other children [54, 55, 56, 57]. The term cyberbullying has recently emerged to denote a new type of bullying that takes place over digital platforms, where the stalkers can perform their crimes on the vulnerable victims. In severe cases, the harassment has lead the victims to the extreme causing irreparable damage or leading them to suicide. In order to stop cyberbullying, the scientific community is developing effective tools capable of detecting the harassment as soon as possible; however, these detection systems are still in an early stage and must be improved. [3] Our contribution is CyberDect, an online-tool that seeks on Social Networks indications of harassment. In a nutshell, our proposal combines Open Source Intelligence tools with Natural Language Processing techniques to analyse posts seeking for abusive language towards the victim. The evaluation of our proposal has been performed with a case-study that consisted in monitor two real high school accounts from Spain.

Jean Piaget and Lev Vygotsky are the two most influential developmental psychologists. Their contributions to the field of developmental psychology, though different, are still similarly remarkable and unique. In spite of such resemblances, there exists a crucial, and generally unnoticed, the difference between Piaget's and Vygotsky's theories, and that this difference underlies the way each author addresses the concept of cognitive development [5]. In short, which theory is more correct? Throughout this paper, we will discover what informs both psychologists' theories, how they are similar, how they are different, and why they have both remained so prominent throughout educational textbooks. Although never in direct competition with each other, the theories developed by Piaget and Vygotsky are often used in contrast with one another, since both offer learning theories with a significant difference, but still impacting on understanding cognitive development.

Workplace cyberbullying (WCB) is a new form of hostility in organizations in which information technology is used as a means to bully employees. The objective of this study is to determine the association between WCB and the interpersonal deviance (ID) [6] of victims through parallel mediation through the ineffectual silence of employees and emotional exhaustion (EE) [6]. Conservation of resource (COR) theory and affective events theory were used as the study's guiding framework, and data were drawn from 351 white-collar employees who were employed in a variety of industries-such as banking, telecommunications sector, education, health care, insurance, and consultancy-in Lahore, Pakistan. The results show that ineffectual silence negatively mediated the relationship between cyberbullying and deviance, decreasing the level of deviance of employees who used silence as a coping mechanism. EE, however, positively mediated the relationship between cyberbullying and deviance. This means that when employees felt emotionally overwhelmed they retaliated by engaging in deviant behaviors and acting as a bully toward colleagues. Drawing on the COR theory and the affective events theory, the findings show that WCB has an impact on ID. From a practical standpoint, the study reveals that WCB can lead to ID and it also may associate with large financial costs and workplace disruptions. Thus, organizations should establish a culture that prevent employees from engaging in WCB and adopt practices of prevention and intervention because it is not only harmful to the employees but also to the organization.

3. METHODOLOGY

i) Proposed Work:

The proposed system leverages advanced techniques, combining machine learning and transfer learning methodologies for robust cyberbullying detection in social networks. Emphasizing nuanced feature extraction and selection, the model aims to enhance contextual understanding. Integrating diverse models, including LinearSVC, Logistic Regression, DistilBert, DistilRoBerta, Electra, the system comprehensively addresses the multifaceted challenge of cyberbullying across various contexts and content formats. As an extension to the project, advanced deep learning models, including LSTM [24] and a hybrid LSTM+GRU architecture, were incorporated alongside a machine learning algorithm, the Voting Classifier. This ensemble method (voting classifier) combined predictions from AdaBoost and RandomForest using a soft voting strategy, enhancing the overall performance of cyberbullying detection. Additionally, a user-friendly Flask framework integrated with SQLite was developed for secure signup and signin, facilitating user testing.

ii) System Architecture:

The cyberbullying detection system follows a systematic process, commencing with the preparation of the cyberbullying dataset [49]. This involves essential data preprocessing steps such as noise removal, normalization, and cleaning. The preprocessed data is then split into training and testing sets. The model building phase employs a comprehensive approach, encompassing traditional machine learning algorithms (LinearSVC, Logistic Regression, and a Voting Classifier combining AdaBoost and Random Forest as an extension to the project), transfer learning models (DistilBert, DistilRoBerta, and Electra), and deep learning architectures (LSTM and LSTM+GRU as extensions to the project again). The trained models are evaluated using a dedicated test set, assessing performance metrics like accuracy, precision, recall, and F1 score for cyberbullying classification. This multifaceted system architecture ensures a robust and versatile solution for cyberbullying detection across diverse contexts and content formats.

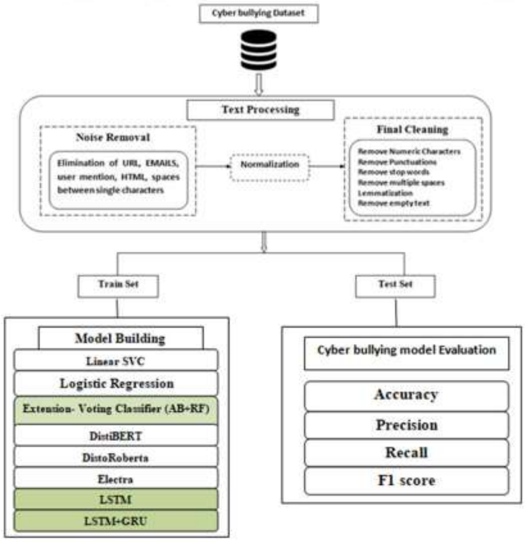
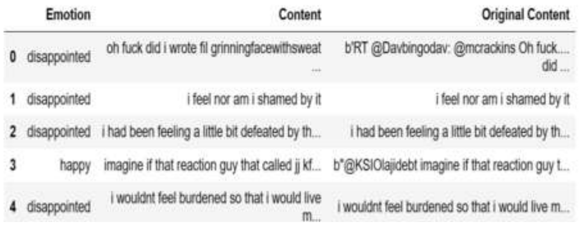


Fig 1 Proposed architecture

iii) Dataset collection:

The cyberbullying dataset [49 is loaded and explored to gain a deeper understanding of its structure, features, and distribution. This initial exploration provides insights into the characteristics of instances related to cyberbullying.



	Emotion	Content	Original Content
0	disappointed	oh fuck did i wrote fil grinningfacewithsweat ...	b'RT @Davbingodav: @mcrackins Oh fuck.... did ...
1	disappointed	i feel nor am i shamed by it	i feel nor am i shamed by it
2	disappointed	i had been feeling a little bit defeated by th...	i had been feeling a little bit defeated by th...
3	happy	imagine if that reaction guy that called jj kf...	b"@KSIOlajidebt imagine if that reaction guy t...
4	disappointed	i wouldnt feel burdened so that i would live m...	i wouldnt feel burdened so that i would live m...

Fig 2 Dataset

iv) Data Processing:

Data processing involves transforming raw data into valuable information for businesses. Generally, data scientists process data, which includes collecting, organizing, cleaning, verifying, analyzing, and converting it into readable formats such as graphs or documents. Data processing can be done using three methods i.e., manual, mechanical, and electronic. The aim is to increase the value of information and facilitate decision-making. This enables businesses to improve their operations and make timely strategic decisions. Automated data processing solutions, such as computer software programming, play a significant role in this. It can help turn large amounts of data, including big data, into meaningful insights for quality management and decision-making.

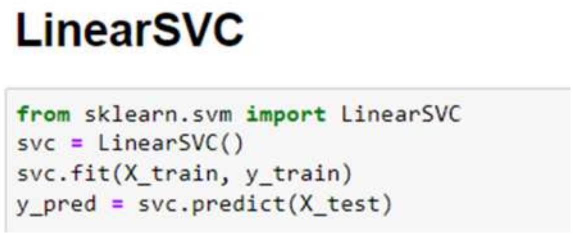
v) Feature selection:

Feature selection is the process of isolating the most consistent, non-redundant, and relevant features to use in model construction. Methodically reducing the size of datasets is important as the size and variety of datasets continue to grow. The main goal of feature selection is to improve the performance of a predictive model and reduce the computational cost of modeling.

Feature selection, one of the main components of feature engineering, is the process of selecting the most important features to input in machine learning algorithms. Feature selection techniques are employed to reduce the number of input variables by eliminating redundant or irrelevant features and narrowing down the set of features to those most relevant to the machine learning model. The main benefits of performing feature selection in advance, rather than letting the machine learning model figure out which features are most important.

vi) Algorithms:

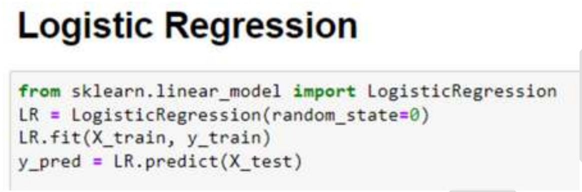
Linear Support Vector Classifier (LinearSVC) is chosen for its effectiveness in binary classification tasks like cyberbullying detection. It works by finding the optimal hyperplane to separate data points, maximizing the margin between classes, making it suitable for distinguishing between bullying and non-bullying content in social media.


LinearSVC

```
from sklearn.svm import LinearSVC
svc = LinearSVC()
svc.fit(X_train, y_train)
y_pred = svc.predict(X_test)
```

Fig 3 LinearSVC

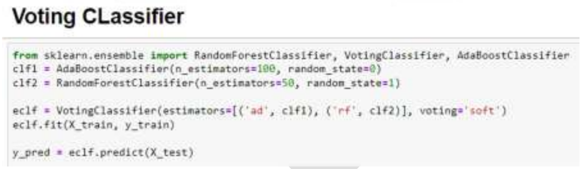
Logistic Regression is employed in the project due to its effectiveness in binary classification tasks, like distinguishing between cyberbullying and non-cyberbullying instances. It models the probability of an instance belonging to a particular class, providing a reliable foundation for discrimination in the system [20,47].


Logistic Regression

```
from sklearn.linear_model import LogisticRegression
LR = LogisticRegression(random_state=0)
LR.fit(X_train, y_train)
y_pred = LR.predict(X_test)
```

Fig 4 Logistic regression

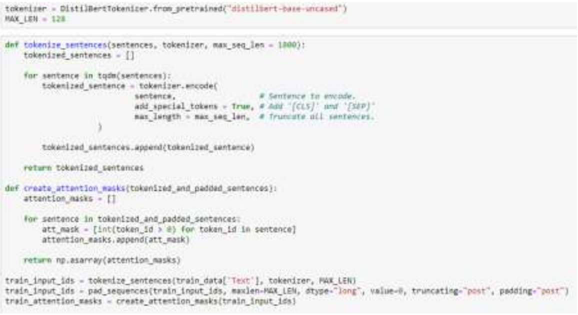
The **Voting Classifier** merges predictions from AdaBoost and RandomForest models through a soft voting strategy, averaging class probabilities. This ensemble technique harnesses diverse algorithms, enhancing the system's resilience and performance in cyberbullying detection across varied social network scenarios. The collaborative decision-making of multiple models contributes to a more comprehensive and accurate classification approach, aligning with the project's goal of addressing the multifaceted challenges of cyberbullying in online interactions.


Voting CLassifier

```
from sklearn.ensemble import RandomForestClassifier, VotingClassifier, AdaBoostClassifier
clf1 = AdaBoostClassifier(n_estimators=100, random_state=0)
clf2 = RandomForestClassifier(n_estimators=50, random_state=1)

eclf = VotingClassifier(estimators=[('ad', clf1), ('rf', clf2)], voting='soft')
eclf.fit(X_train, y_train)

y_pred = eclf.predict(X_test)
```

Fig 5 Voting classifier

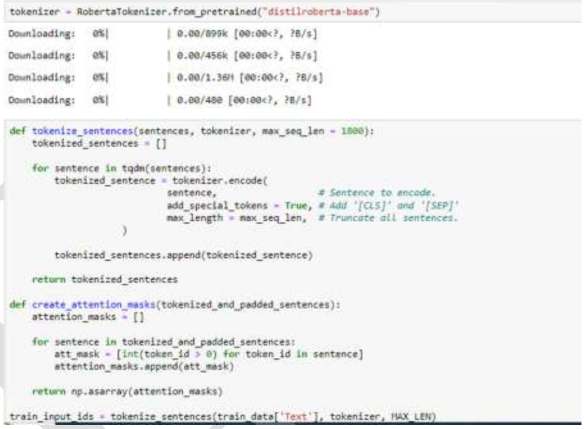
DistilBERT, a distilled version of BERT, retains contextual language understanding with fewer parameters. It achieves computational efficiency by removing unnecessary components, making it ideal for cyberbullying detection where a balance between model sophistication and resource efficiency is crucial for processing large volumes of social media data [53, 54, 55].



```
tokenizer = DistilBertTokenizer.from_pretrained("distilbert-base-uncased")
MAX_LEN = 128

def tokenize_sentences(sentences, tokenizer, max_seq_len = 1000):
    tokenized_sentences = []

    for sentence in tqdm(sentences):
        tokenized_sentence = tokenizer.encode(
                            sentence,                      # Sentence to encode.
                            add_special_tokens = True, # Add '[CLS]' and '[SEP]'
                            max_length = max_seq_len,  # Truncate all sentences.
                    )

        tokenized_sentences.append(tokenized_sentence)

    return tokenized_sentences

def create_attention_masks(tokenized_and_padded_sentences):
    attention_masks = []

    for sentence in tokenized_and_padded_sentences:
        att_mask = [int(token_id > 0) for token_id in sentence]
        attention_masks.append(att_mask)

    return np.asarray(attention_masks)

train_input_ids = tokenize_sentences(train_data['Text'], tokenizer, MAX_LEN)
train_input_ids = pad_sequences(train_input_ids, maxlen=MAX_LEN, dtype="long", value=0, truncating="post", padding="post")
train_attention_masks = create_attention_masks(train_input_ids)
```

Fig 6 DistilBERT

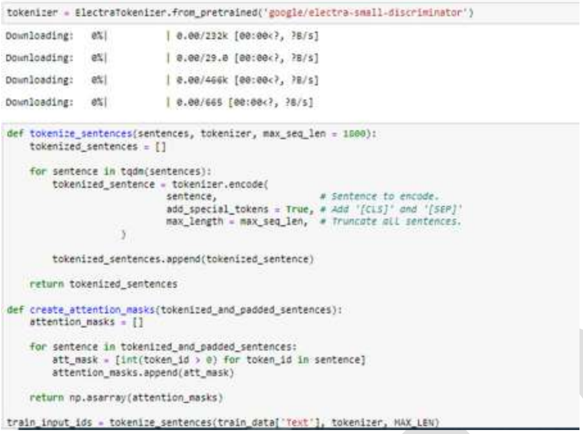
DistilRoBERTa is a distilled version of RoBERTa [53, 54, 55], a robust transformer-based language model. It leverages knowledge distillation to retain the essential linguistic understanding of RoBERTa in a more compact form. DistilRoBERTa is chosen for this project due to its efficient representation learning, striking a balance between performance and computational resources in cyberbullying detection tasks.



```
tokenizer = RobertaTokenizer.from_pretrained("distilroberta-base")

Downloading:   0%|          | 0.00/899k [00:00<?, ?B/s]

Downloading:   0%|          | 0.00/456k [00:00<?, ?B/s]

Downloading:   0%|          | 0.00/1.36M [00:00<?, ?B/s]

Downloading:   0%|          | 0.00/480 [00:00<?, ?B/s]

def tokenize_sentences(sentences, tokenizer, max_seq_len = 1000):
    tokenized_sentences = []

    for sentence in tqdm(sentences):
        tokenized_sentence = tokenizer.encode(
                            sentence,                      # Sentence to encode.
                            add_special_tokens = True, # Add '[CLS]' and '[SEP]'
                            max_length = max_seq_len,  # Truncate all sentences.
                    )

        tokenized_sentences.append(tokenized_sentence)

    return tokenized_sentences

def create_attention_masks(tokenized_and_padded_sentences):
    attention_masks = []

    for sentence in tokenized_and_padded_sentences:
        att_mask = [int(token_id > 0) for token_id in sentence]
        attention_masks.append(att_mask)

    return np.asarray(attention_masks)

train_input_ids = tokenize_sentences(train_data['Text'], tokenizer, MAX_LEN)
```

Fig 7 DistilRoBERTa

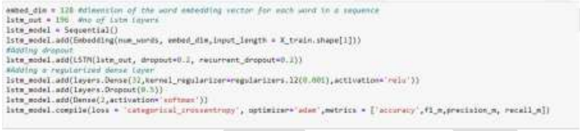
Electra is a pre-trained language model that replaces a portion of input text with incorrect words and trains the model to discern genuine from altered content. It enhances the proposed system's robustness by improving the model's understanding of subtle nuances and variations in language, crucial for accurate cyberbullying detection in diverse social network contexts.



```
tokenizer = ElectraTokenizer.from_pretrained('google/electra-small-discriminator')

Downloading:   0%|          | 0.00/232k [00:00<?, ?B/s]

Downloading:   0%|          | 0.00/29.0 [00:00<?, ?B/s]

Downloading:   0%|          | 0.00/466k [00:00<?, ?B/s]

Downloading:   0%|          | 0.00/665 [00:00<?, ?B/s]

def tokenize_sentences(sentences, tokenizer, max_seq_len = 1000):
    tokenized_sentences = []

    for sentence in tqdm(sentences):
        tokenized_sentence = tokenizer.encode(
                            sentence,                  # Sentence to encode.
                            add_special_tokens = True, # Add '[CLS]' and '[SEP]'
                            max_length = max_seq_len,  # Truncate all sentences.
                    )

        tokenized_sentences.append(tokenized_sentence)

    return tokenized_sentences

def create_attention_masks(tokenized_and_padded_sentences):
    attention_masks = []

    for sentence in tokenized_and_padded_sentences:
        att_mask = [int(token_id > 0) for token_id in sentence]
        attention_masks.append(att_mask)

    return np.asarray(attention_masks)

train_input_ids = tokenize_sentences(train_data['Text'], tokenizer, MAX_LEN)
```

Fig 8 Electra

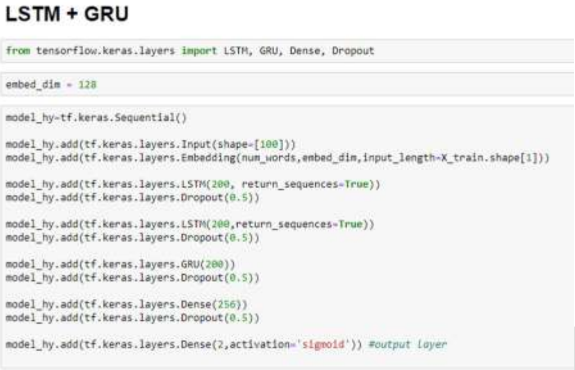
Long Short-Term Memory (LSTM) is a specialized recurrent neural network (RNN) variant employed in this project to analyze social media interactions. LSTMs excel at learning intricate patterns over extended sequences by employing memory cells and gating mechanisms. This design enables the model to effectively capture and remember long-term dependencies, making LSTMs well-suited for the nuanced task of cyberbullying detection, where understanding complex contextual relationships is crucial for accurate classification.



```
embed_dim = 128 #dimension of the word embedding vector for each word in a sequence
lstm_out = 196  #no of lstm layers
lstm_model = Sequential()
lstm_model.add(Embedding(num_words, embed_dim,input_length = X_train.shape[1]))
#Adding dropout
lstm_model.add(LSTM(lstm_out, dropout=0.2, recurrent_dropout=0.2))
#Adding a regularized dense layer
lstm_model.add(layers.Dense(32,kernel_regularizer=regularizers.l2(0.001),activation='relu'))
lstm_model.add(layers.Dropout(0.5))
lstm_model.add(Dense(2,activation='softmax'))
lstm_model.compile(loss = 'categorical_crossentropy', optimizer='adam',metrics = ['accuracy',f1_m,precision_m, recall_m])
```

Fig 9 LSTM

LSTM (Long Short-Term Memory) and GRU (Gated Recurrent Unit) are specialized types of recurrent neural networks. By combining LSTM and GRU, the model benefits from improved memory retention and processing efficiency, essential for comprehending the intricacies of sequential data like social media interactions. This hybrid architecture enhances the system's capability to recognize and interpret long-term dependencies, providing a more effective solution for cyberbullying detection in the dynamic and nuanced context of social networks [47, 48].



LSTM + GRU

```
from tensorflow.keras.layers import LSTM, GRU, Dense, Dropout
```

```
embed_dim = 128
```

```
model_hy=tf.keras.Sequential()

model_hy.add(tf.keras.layers.Input(shape=[100]))
model_hy.add(tf.keras.layers.Embedding(num_words,embed_dim,input_length=X_train.shape[1]))

model_hy.add(tf.keras.layers.LSTM(200, return_sequences=True))
model_hy.add(tf.keras.layers.Dropout(0.5))

model_hy.add(tf.keras.layers.LSTM(200,return_sequences=True))
model_hy.add(tf.keras.layers.Dropout(0.5))

model_hy.add(tf.keras.layers.GRU(200))
model_hy.add(tf.keras.layers.Dropout(0.5))

model_hy.add(tf.keras.layers.Dense(256))
model_hy.add(tf.keras.layers.Dropout(0.5))

model_hy.add(tf.keras.layers.Dense(2,activation='sigmoid')) #output layer
```

Fig 10 LSTM + GRU

4. EXPERIMENTAL RESULTS

Precision: Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\text{Precision} = \frac{\textit{True Positive}}{\textit{True Positive}+\textit{False Positive}}$$

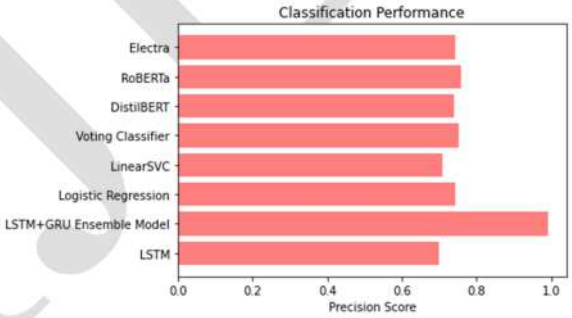


Fig 11 Precision comparison graph

Recall: Recall is a metric in machine learning that measures the ability of a model to identify all relevant instances of a particular class. It is the ratio of correctly predicted positive observations to the total actual positives, providing insights into a model's completeness in capturing instances of a given class.

$$Recall = \frac{TP}{TP + FN}$$

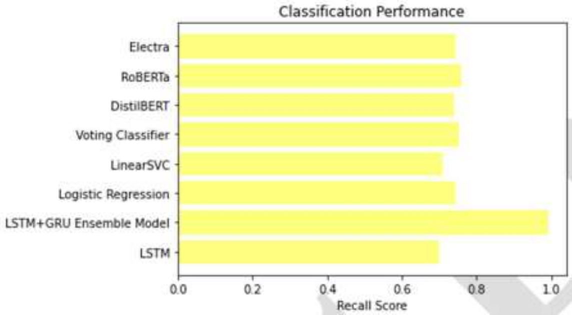


Fig 12 Recall comparison graph

Accuracy: Accuracy is the proportion of correct predictions in a classification task, measuring the overall correctness of a model's predictions.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

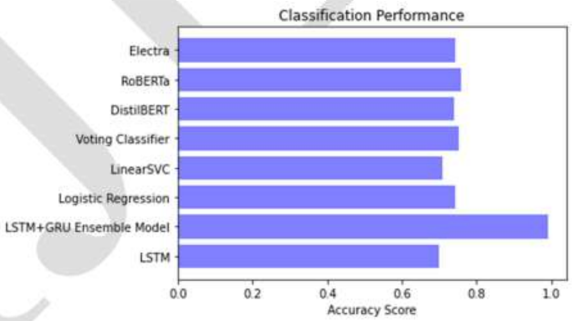


Fig 13 Accuracy graph

F1 Score: The F1 Score is the harmonic mean of precision and recall, offering a balanced measure that considers both false positives and false negatives, making it suitable for imbalanced datasets.

$$\text{F1 Score} = 2 * \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} * 100$$

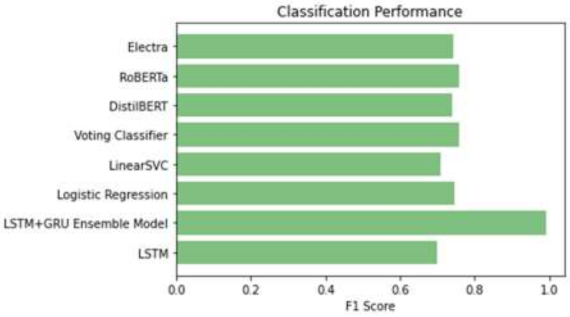
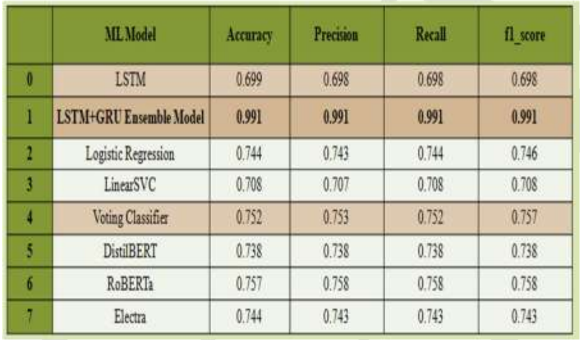
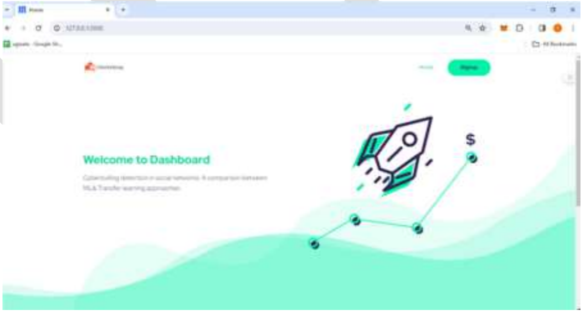


Fig 14 F1Score



	ML Model	Accuracy	Precision	Recall	f1_score
0	LSTM	0.699	0.698	0.698	0.698
1	**LSTM+GRU Ensemble Model**	**0.991**	**0.991**	**0.991**	**0.991**
2	Logistic Regression	0.744	0.743	0.744	0.746
3	LinearSVC	0.708	0.707	0.708	0.708
4	Voting Classifier	0.752	0.753	0.752	0.757
5	DistilBERT	0.738	0.738	0.738	0.738
6	RoBERTa	0.757	0.758	0.758	0.758
7	Electra	0.744	0.743	0.743	0.743

Fig 15 Performance Evaluation



Welcome to Dashboard

Cyberbulling detection in social networks: A comparison between ML & Transfer learning approaches

Fig 16 Home page

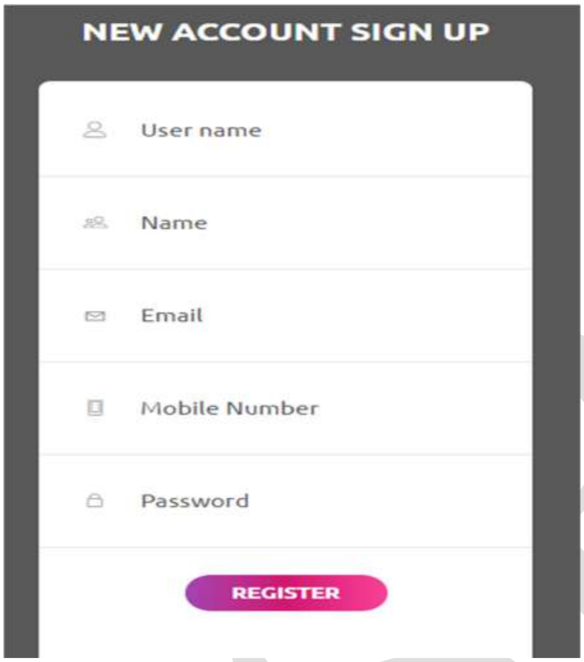
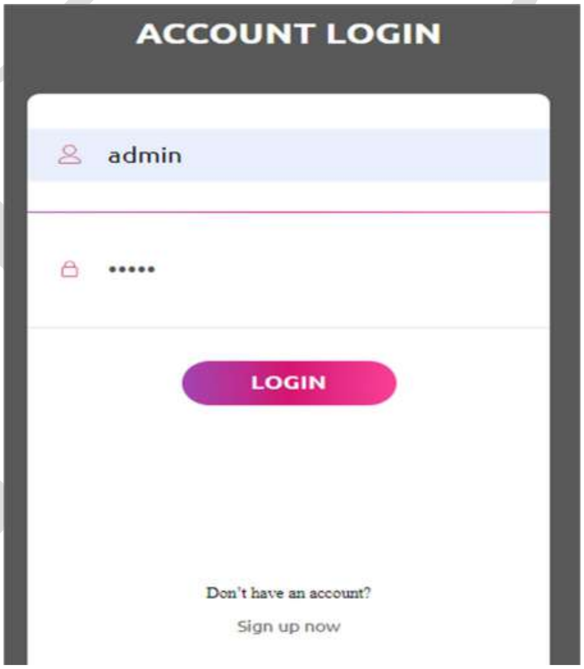


Fig 17 Signin page

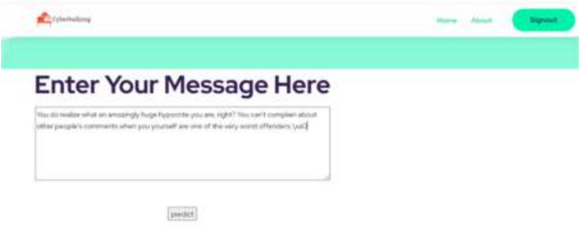


admin

•••••

Don't have an account?

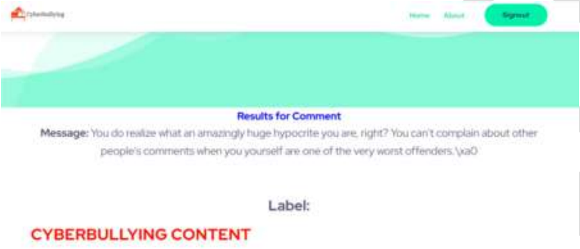
Fig 18 Login page



Enter Your Message Here

You do realize what an amazingly huge hypocrite you are, right? You can't complain about other people's comments when you yourself are one of the very worst offenders.\xa0

Fig 19 User input



Results for Comment

Message: You do realize what an amazingly huge hypocrite you are, right? You can't complain about other people's comments when you yourself are one of the very worst offenders.\xa0

Label:

CYBERBULLYING CONTENT

Fig 20 Predict result for given input

5. CONCLUSION

The project successfully addresses the pressing issue of cyberbullying by employing a diverse set of machine learning algorithms, including LinearSVC, Logistic Regression, DistilBert, DistilRoBerta, and Electra. This ensures a robust approach to identify instances of cyberbullying in online content. Leveraging advanced natural language processing techniques, the project incorporates features such as text normalization, tokenization, and bag-of-words representation. This enables a nuanced understanding of the language used in online posts, contributing to the accurate detection of cyberbullying. The project involves the implementation of Long Short-Term Memory (LSTM), LSTM+GRU and a voting classifier combining predictions from multiple individual models [47,48]. LSTM+GRU outperformed all other models. This innovative approach enhances the accuracy and reliability of cyberbullying predictions, providing a more robust final outcome. The development of a front-end using the Flask framework ensures a user-friendly experience for individuals interacting with the system. Incorporating user authentication through SQLite adds an additional layer of security, creating a reliable platform for users to engage with the cyberbullying detection system. The front-end design allows for user testing, input validation, and seamless model predictions, enhancing practical usability. By applying Latent Dirichlet Allocation (LDA) [33, 37]for topic modeling, the project goes beyond simple classification, offering valuable insights into underlying themes in cyberbullying content. This contributes to a comprehensive understanding of the issue and provides additional context for addressing and preventing cyberbullying in online communities.

6. FUTURE SCOPE

The future scope involves expanding the system to include the analysis of diverse content formats, such as images, videos, and audio. This enhancement aims to achieve a more comprehensive understanding of cyberbullying across various modalities. Future development will focus on improving the system for real-time cyberbullying detection [23, 47]. This will enable timely intervention and support, contributing to a safer online environment for users. The project will implement mechanisms for continuous learning and model adaptation. This approach ensures that the system remains effective against evolving cyberbullying behaviors, providing sustained protection. To broaden its applicability, the project will incorporate multilingual capabilities. This expansion aims to enable the detection of cyberbullying across diverse linguistic contexts, making the solution more inclusive and globally relevant.

REFERENCES

[1] B. Cagirkan and G. Bilek, ‘‘Cyberbullying among Turkish high school students,’’ Scandin. J. Psychol., vol. 62, no. 4, pp. 608–616, Aug. 2021, doi: 10.1111/sjop.12720.

[2] P. T. L. Chi, V. T. H. Lan, N. H. Ngan, and N. T. Linh, ‘‘Online time, experience of cyber bullying and practices to cope with it among high school students in Hanoi,’’ Health Psychol. Open, vol. 7, no. 1, Jan. 2020, Art. no. 205510292093574, doi: 10.1177/2055102920935747.

[3] A. López-Martínez, J. A. García-Díaz, R. Valencia-García, and A. Ruiz-Martínez, ‘‘CyberDect. A novel approach for cyberbullying detection on Twitter,’’ in Proc. Int. Conf. Technol. Innov., Guayaquil, Ecuador: Springer, 2019, pp. 109–121, doi: 10.1007/978-3-030-34989- 9_9.

[4] R. M. Kowalski and S. P. Limber, ‘‘Psychological, physical, and academic correlates of cyberbullying and traditional bullying,’’ J. Adolescent Health, vol. 53, no. 1, pp. S13–S20, Jul. 2013, doi: 10.1016/j.jadohealth.2012.09.018.

[5] Y.-C. Huang, ‘‘Comparison and contrast of piaget and Vygotsky’s theories,’’ in Proc. Adv. Social Sci., Educ. Humanities Res., 2021, pp. 28–32, doi: 10.2991/assehr.k.210519.007.

[6] A. Anwar, D. M. H. Kee, and A. Ahmed, ‘‘Workplace cyberbullying and interpersonal deviance: Understanding the mediating effect of silence and emotional exhaustion,’’ Cyberpsychol., Behav., Social Netw., vol. 23, no. 5, pp. 290–296, May 2020, doi: 10.1089/cyber.2019.0407.

[7] D. M. H. Kee, M. A. L. Al-Anesi, and S. A. L. Al-Anesi, ‘‘Cyberbullying on social media under the influence of COVID-19,’’ Global Bus. Organizational Excellence, vol. 41, no. 6, pp. 11–22, Sep. 2022, doi: 10.1002/joe.22175.

[8] I. Kwan, K. Dickson, M. Richardson, W. MacDowall, H. Burchett, C. Stansfield, G. Brunton, K. Sutcliffe, and J. Thomas, ‘‘Cyberbullying and children and young people’s mental health: A systematic map of systematic reviews,’’ Cyberpsychol., Behav., Social Netw., vol. 23, no. 2, pp. 72–82, Feb. 2020, doi: 10.1089/cyber.2019.0370.

[9] R. Garett, L. R. Lord, and S. D. Young, ‘‘Associations between social media and cyberbullying: A review of the literature,’’ mHealth, vol. 2, p. 46, Dec. 2016, doi: 10.21037/mhealth.2016.12.01.

[10] M. Ptaszynski, F. Masui, Y. Kimura, R. Rzepka, and K. Araki, ‘‘Automatic extraction of harmful sentence patterns with application in cyberbullying detection,’’ in Proc. Lang. Technol. Conf. Poznań, Poland: Springer, 2015, pp. 349–362, doi: 10.1007/978-3-319-93782-3_25.

[11] M. Ptaszynski, P. Lempa, F. Masui, Y. Kimura, R. Rzepka, K. Araki, M. Wroczynski, and G. Leliwa, ‘‘"Brute-force sentence pattern extortion from harmful messages for cyberbullying detection,’’’ J. Assoc. Inf. Syst., vol. 20, no. 8, pp. 1075–1127, 2019.

[12] M. O. Raza, M. Memon, S. Bhatti, and R. Bux, ‘‘Detecting cyberbullying in social commentary using supervised machine learning,’’ in Proc. Future Inf. Commun. Conf. Cham, Switzerland: Springer, 2020, pp. 621–630.

[13] D. Nguyen, M. Liakata, S. Dedeo, J. Eisenstein, D. Mimno, R. Tromble, and J. Winters, ‘‘How we do things with words: Analyzing text as social and cultural data,’’ Frontiers Artif. Intell., vol. 3, p. 62, Aug. 2020, doi: 10.3389/frai.2020.00062.

[14] J. Cai, J. Li, W. Li, and J. Wang, ‘‘Deeplearning model used in text classification,’’ in Proc. 15th Int. Comput. Conf. Wavelet Act. Media Technol. Inf. Process. (ICCWAMTIP), Dec. 2018, pp. 123–126, doi: 10.1109/ICCWAMTIP.2018.8632592.

[15] N. Tiku and C. Newton. Twitter CEO: We Suck at Dealing With Abuse. Verge. Accessed: Aug. 17, 2022. [Online]. Available: https://www.theverge.com/2015/2/4/7982099/twitter-ceo-sent-memotaking-personal-responsibility-for-the [16] D. Noever, ‘‘Machine learning suites for online toxicity detection,’’ 2018, arXiv:1810.01869.

[17] D. G. Krutka, S. Manca, S. M. Galvin, C. Greenhow, M. J. Koehler, and E. Askari, ‘‘Teaching ‘against’ social media: Confronting problems of profit in the curriculum,’’ Teachers College Rec., Voice Scholarship Educ., vol. 121, no. 14, pp. 1–42, Dec. 2019, doi: 10.1177/016146811912101410.

[18] H. Rosa, N. Pereira, R. Ribeiro, P. C. Ferreira, J. P. Carvalho, S. Oliveira, L. Coheur, P. Paulino, A. M. V. Simão, and I. Trancoso, ‘‘Automatic cyberbullying detection: A systematic review,’’ Comput. Hum. Behav., vol. 93, pp. 333–345, Apr. 2019, doi: 10.1016/j.chb.2018.12.021.

[19] S. Bharti, A. K. Yadav, M. Kumar, and D. Yadav, ‘‘Cyberbullying detection from tweets using deep learning,’’ Kybernetes, vol. 51, no. 9, pp. 2695–2711, Sep. 2022.

[20] A. Bozyiğit, S. Utku, and E. Nasibov, ‘‘Cyberbullying detection: Utilizing social media features,’’ Expert Syst. Appl., vol. 179, Oct. 2021, Art. no. 115001, doi: 10.1016/j.eswa.2021.115001.

[21] H.-S. Lee, H.-R. Lee, J.-U. Park, and Y.-S. Han, ‘‘An abusive text detection system based on enhanced abusive and non-abusive word lists,’’ Decis. Support Syst., vol. 113, pp. 22–31, Sep. 2018, doi: 10.1016/j.dss.2018.06.009.

[22] Y. Fang, S. Yang, B. Zhao, and C. Huang, ‘‘Cyberbullying detection in social networks using bi-GRU with self-attention mechanism,’’ Information, vol. 12, no. 4, p. 171, Apr. 2021, doi: 10.3390/info12040171.

[23] G. Jacobs, C. Van Hee, and V. Hoste, ‘‘Automatic classification of participant roles in cyberbullying: Can we detect victims, bullies, and bystanders in social media text?’’ Natural Lang. Eng., vol. 28, no. 2, pp. 141–166, Mar. 2022, doi: 10.1017/S135132492000056X.

[24] M. Gada, K. Damania, and S. Sankhe, ‘‘Cyberbullying detection using LSTM-CNN architecture and its applications,’’ in Proc. Int. Conf. Comput. Commun. Informat. (ICCCI), Jan. 2021, pp. 1–6, doi: 10.1109/ICCCI50826.2021.9402412.

[25] H. H.-P. Vo, H. Trung Tran, and S. T. Luu, ‘‘Automatically detecting cyberbullying comments on online game forums,’’ in Proc. RIVF Int. Conf. Comput. Commun. Technol. (RIVF), Aug. 2021, pp. 1–5, doi: 10.1109/RIVF51545.2021.9642116.

[26] F. Elsafoury, S. Katsigiannis, Z. Pervez, and N. Ramzan, ‘‘When the timeline meets the pipeline: A survey on automated cyberbullying detection,’’ IEEE Access, vol. 9, pp. 103541–103563, 2021, doi: 10.1109/ACCESS.2021.3098979.

[27] J. Howard and S. Ruder, ‘‘Universal language model fine-tuning for text classification,’’ in Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (Long Papers), vol. 1, 2018, pp. 328–339.

[28] R. Silva Barbon and A. T. Akabane, ‘‘Towards transfer learning techniques—BERT, DistilBERT, BERTimbau, and DistilBERTimbau for automatic text classification from different languages: A case study,’’ Sensors, vol. 22, no. 21, p. 8184, Oct. 2022, doi: 10.3390/s22218184.

[29] J. Eronen, M. Ptaszynski, F. Masui, G. Leliwa, and M. Wroczynski, ‘‘Exploring the potential of feature density in estimating machine learning classifier performance with application to cyberbullying detection,’’ 2022, arXiv:2206.01949.

[30] J. Bhagya and P. S. Deepthi, Cyberbullying Detection on Social Media Using SVM (Inventive Systems and Control). Singapore: Springer, 2021, pp. 17–27, doi: 10.1007/978-981-16-1395-1_2.

[31] A. Perera and P. Fernando, ''Accurate cyberbullying detection and prevention on social media,'' Proc. Comput. Sci., vol. 181, pp. 605–611, Jan. 2021, doi: 10.1016/j.procs.2021.01.207.

[32] 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, doi: 10.1145/2833312.2849567.

[33] V. Nahar, X. Li, and C. Pang, ''An effective approach for cyberbullying detection,'' Commun. Inf. Sci. Manage. Eng., vol. 3, no. 5, p. 238, 2013.

[34] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, ''Common sense reasoning for detection, prevention, and mitigation of cyberbullying,'' ACM Trans. Interact. Intell. Syst., vol. 2, no. 3, pp. 1–30, Sep. 2012.

[35] K. Dinakar, R. Reichart, and H. Lieberman, ''Modeling the detection of textual cyberbullying,'' in Proc. Int. AAAI Conf. Web Social Media, Barcelona, Spain, 2011, pp. 11–17. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/14209

[36] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, ''Detection of harassment on web 2.0,'' in Proc. Content Anal. Web, vol. 2. Madrid, Spain, 2009, pp. 1–7. [Online]. Available: https://www.academia.edu/download/47631616/Yin_etal_CAW2009.pdf

[37] C. Van Hee, G. Jacobs, C. Emmery, B. Desmet, E. Lefever, B. Verhoeven, G. De Pauw, W. Daelemans, and V. Hoste, ''Automatic detection of cyberbullying in social media text,'' PLoS ONE, vol. 13, no. 10, Oct. 2018, Art. no. e0203794, doi: 10.1371/journal.pone.0203794.

[38] C. Van Hee, ''Detection and fine-grained classification of cyberbullying events,'' in Proc. Int. Conf. Recent Adv. Natural Lang. Process. (RANLP), 2015, pp. 672–680.

[39] A. Muneer and S. M. Fati, ''A comparative analysis of machine learning techniques for cyberbullying detection on Twitter,'' Future Internet, vol. 12, no. 11, p. 187, Oct. 2020, doi: 10.3390/fi12110187.

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

[41] D. Chatzakou, I. Leontiadis, J. Blackburn, E. D. Cristofaro, G. Stringhini, A. Vakali, and N. Kourtellis, ''Detecting cyberbullying and cyberaggression in social media,'' ACM Trans. Web, vol. 13, no. 3, pp. 1–51, Aug. 2019.

[42] V. Balakrishnan, S. Khan, T. Fernandez, and H. R. Arabnia, ''Cyberbullying detection on Twitter using big five and dark triad features,'' Personality Individual Differences, vol. 141, pp. 252–257, Apr. 2019, doi: 10.1016/j.paid.2019.01.024.

[43] R. Zhao and K. Mao, ‘‘Cyberbullying detection based on semanticenhanced marginalized denoising auto-encoder,’’ IEEE Trans. Affect. Comput., vol. 8, no. 3, pp. 328–339, Jul. 2017, doi: 10.1109/TAFFC. 2016.2531682.

[44] X. Zhang, J. Tong, N. Vishwamitra, E. Whittaker, J. P. Mazer, R. Kowalski, H. Hu, F. Luo, J. Macbeth, and E. Dillon, ‘‘Cyberbullying detection with a pronunciation based convolutional neural network,’’ in Proc. 15th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA), Dec. 2016, pp. 740–745, doi: 10.1109/ICMLA.2016.0132.

[45] A. Kumar and N. Sachdeva, ‘‘A bi-GRU with attention and CapsNet hybrid model for cyberbullying detection on social media,’’ World Wide Web, vol. 25, no. 4, pp. 1537–1550, Jul. 2022, doi: 10.1007/s11280-021- 00920-4.

[46] K. Shriniket, P. Vidyarthi, S. Udyavara, R. Manohar, and N. Shruthi, ‘‘A time optimised model for cyberbullying detection,’’ Int. Res. J. Modernization Eng., Technol. Sci., vol. 4, no. 7, pp. 808–815, 2022.

[47] S. Agrawal and A. Awekar, ‘‘Deep learning for detecting cyberbullying across multiple social media platforms,’’ in Proc. Eur. Conf. Inf. Retr. (ECIR), Grenoble, France: Springer, 2018, pp. 141–153, doi: 10.1007/978-3-319-76941-7_11.

[48] M. Dadvar and K. Eckert, ‘‘Cyberbullying detection in social networks using deep learning based models; a reproducibility study,’’ 2018, arXiv:1812.08046.

[49] H. Rosa, D. Matos, R. Ribeiro, L. Coheur, and J. P. Carvalho, ‘‘A ‘deeper’ look at detecting cyberbullying in social networks,’’ in Proc. Int. Joint Conf. Neural Netw. (IJCNN), Jul. 2018, pp. 1–8, doi: 10.1109/IJCNN.2018.8489211.

[50] 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, Calgary, ALB, Canada: SIAM, 2019, pp. 235–243.