Industrial Engineering Journal

ISSN: 0970-2555

Volume : 53, Special Issue 4, April : 2024

CYBERBULLYING DETECTION BASED ON EMOTION

Mr.D KarunaKumar Reddy1,Y J Sravanthi2, CH S C Pravalika3,D Yasawini4,A Bhavya Sri5
1,2,3,4,5 Department of Computer Science and Engineering, Vignan's Institute of Engineering for Women,Visakhapatnam,Andhra Pradesh,India

ABSTRACT

Due to the adverse impacts of cyberbullying, a great deal of research has been conducted to identify effective solutions to this reoccurring problem. This research is motivated by the fact that cyberbullying may cause negative emotions. We proposes a cyberbullying detection models that are trained based on contextual, emotions and sentiment features. An Emotion Detection Model (EDM) was constructed using Twitter datasets that have been improved in terms of its annotations. Emotions and sentiment were extracted from cyberbullying datasets using EDM and lexicons based. Two cyberbullying datasets such as Toxic dataset is collected from the Wikipedia and Twitter data set respectively were further improved by comprehensive annotation of emotion and sentiment features. The results show that anger, fear and guilt were the major emotions associated with cyberbullying. Subsequently, the extracted emotions were used as features in addition to contextual and sentiment features to train models for cyberbullying detection. The results demonstrate that using emotion features and sentiment has improved the performance of detecting cyberbullying. The proposed models also outperformed the state-of-the-art models with good f1-score. The main contribution of this work is two-fold, which includes a comprehensive emotion annotated dataset for cyberbullying detection, and an empirical proof of emotions as effective features for cyberbullying detection.

Keywords:Cyberbullying, Emotion-based detection, Natural language processing, Sentiment analysis, Textual analysis, Online communities, Cybersecurity, False positive reduction, Context-aware detection

INTRODUCTION

Theadvancement of information and communication technologies has provided an avenue for the online community to publish and respond to user-generated content (UGC). Unfortunately, such convenience has been abused by online bullies, causing harm to others via threatening, harassing, humiliating, intimidating, manipulating, or controlling targeted victims. These acts, termed 'cyberbullying,' are defined as the willful and repeated harm inflicted using electronic devices. Cyberbullying (CB) can have a severe impact on a victim's mental health, ranging from negative emotions (anger, fear, sadness, guilt, etc.) to depression, and even suicidal thoughts. Emotion mining is one of the focus areas in affective computing that aims to identify, analyze, and evaluate the human state of mind towards various events or encounters. Emotion analysis has significantly impacted different sectors such as the stock market, consumers' feedback, and recommendations. However, researchers have not deeply taken emotion analysis into consideration for cyberbullying detection. Therefore, including emotion features to facilitate cyberbullying detection can improve detection accuracy due to the strong relationship between cyberbullying and negative emotions.

LITERATURESURVEY

Leveraging semantic features for recommendation: Sentence-level emotion analysis:

Personalized recommendation systems can help users to filter redundant information from a large amount of data. Previous relevant researches focused on learning user preferences by analyzing texts from comment communities without exploring the detailed sentiment polarity, which encountered the cold-start problem .To address this research gap, we propose a hybrid personalized recommendation model that extracts user preferences by analyzing user review content in different sentiment polarity at the sentence level, based on jointly applying user-item score matrices and dimension reduction method ..A novel voting mechanism is also designed based on positive preferences from the neighbors of the target user to directly generate the recommendation results.

The experimental results of testing the proposed model with a real-world data set show that our proposed model can achieve better recommendation effects than the representative recommendation algorithms. In addition, we demonstrated that fine-grained emotion recognition has good adaptability to a sparse rating matrix with a reasonable and good performance.

Detecting the target of sarcasm is hard: Really??:

Sarcasm target detection (identifying the target of mockery in a sarcastic sentence) is an emerging field in computational linguistics. Although there has been some research in this field, accurately identifying the target still remains problematic especially when the target of mockery is not presented in the text. In this paper, we propose a combination of a machine learning classifier and a deep learning model to extract the target of sarcasm from the text. First, we classify sarcastic sentences using machine learning, to determine whether a sarcastic sentence contains a target. Then we use a deep learning model from Aspect-Based Sentiment Analysis to extract the target. Our proposed system is evaluated on three publicly available data sets: sarcastic book snippets, sarcastic tweets, and sarcastic Reddit comments. Our evaluation results show that our approach achieves equal or better performance compared to the current state-of-the-art system, with an 18% improvement on the Reddit data set and similar scores on the Books and Tweets data sets. This is because our method is able to accurately identify when the target of sarcasm is not present. The primary challenge we identify, that is hindering the creation of a high accuracy classifier, is the lack of consistency among human annotators in identifying the target of sarcasm within standard ground-truth data sets.

Machine Learning and feature engineering-based study into sarcasm and irony classification with application to cyberbullying detection:Irony and sarcasm detection is considered a complex task in Natural

Language Processing. This paper set out to explore the sarcasm and irony on Twitter, using Machine Learning and Feature Engineering techniques. First we review and clarify the definition of irony and sarcasm by discussing various studies focusing on the terms. Next the first experiment is conducted comparing between various types of classification methods including some popular classifiers for text classification task. For the second experiment, different types of data preprocessing methods were compared and analyzed. Finally, the relationship between irony, sarcasm, and cyberbullying are discussed. The results are interesting as we observed high similarity between them.

Improving classifier training efficiency for automatic cyberbullying detection with Feature Density:
We study the effectiveness of Feature Density (FD) using different linguistically-backed feature preprocessing methods in order to estimate dataset complexity, which in turn is used to comparatively estimate the potential performance of machine learning (ML) classifiers prior to any training. We hypothesize that estimating dataset complexity allows for the reduction of the number of required experiments iterations. This way we can optimize the resource-intensive training of ML models which is becoming a serious issue due to the increases in available dataset sizes and the ever rising popularity of models based on Deep Neural Networks (DNN). The problem of constantly increasing needs for more powerful computational resources is also affecting the environment due to alarmingly-growing amount of CO2 emissions caused by training of large-scale ML models. The research was conducted on multiple datasets, including popular datasets, such as Yelp business review dataset used for training typical sentiment analysis models, as well as more recent datasets trying to tackle the problem of cyberbullying, which, being a serious social problem, is also a much more sophisticated problem form the point of view of linguistic representation. We use cyberbullying datasets collected for multiple languages, namely English, Japanese and Polish. The difference in linguistic complexity of datasets allows us to additionally discuss the efficacy of linguistically-backed word preprocessing.

Detection of Harassment Type of Cyberbullying: A Dictionary of Approach Words and Its Impact: The purpose of this paper is to analyse the effects of predatory approach words in the detection of cyberbullying and to propose a mechanism of generating a dictionary of such approach words. The research incorporates analysis of chat logs from convicted felons, to generate a dictionary of sexual approach words. By analysing data across multiple social networks, the study demonstrates the usefulness of such a dictionary

of approach words in detection of online predatory behaviour through machine learning algorithms.It also shows the difference between the nature of contents across specific social network platforms.This research is tailored to focus on sexual harassment type of cyberbullying and proposes a novel dictionary of approach words. Since cyberbullying is a growing threat to the mental health and intellectual development of adolescents in the society, models targeted towards the detection of specific type of online bullying or predation should be encouraged among social network researchers.

EXISTINGSYSTEM

In literature they understand the characteristics of abusive behavior in Twitter, one of the largest social media platforms. They analyze 1.2 million users and 2.1 million tweets, comparing users participating in discussions around seemingly normal topics like the NBA, to those more likely to be hate-related, such as the Gamergate controversy, or the gender pay inequality at the BBC station. They also explore specific manifestations of abusive behavior, i.e., cyberbullying and cyber aggression, in one of the hate-related communities (Gamergate). They present a robust methodology to distinguish bullies and aggressors from normal Twitter users by considering text, user, and network-based attributes. They use various state-of-the- art machine-learning algorithms to classify these accounts.

Disadvantages

The existing work primarily analyzes abusive behavior on Twitter with a focus on distinguishing bullies and aggressors from normal users based on text, user, and network-based attributes.
The existing work does not seem to explicitly address emotion detection, which can be a crucial aspect of understanding abusive behavior.
The existing work's reliance on machine learning algorithms for classification. This approach may result in less accuracy in cyberbullying detection.
The existing work's focus on text, user, and network-based attributes might miss out on the rich emotional context that can be provided by the present work's emphasis on emotions and sentiments.

PROPOSEDSCHEME

We proposes a cyberbullying detection models (CBMs) that are trained based on contextual, emotions and sentiment features. An Emotion Detection Model (EDM) was constructed using Twitter datasets that have been improved in terms of its annotations. In our work we use mainly the two types of data sets such as cyberbullying data sets and emotion datasets. The cyberbullying datasets such as Toxic dataset is collected from the Wikipedia and Twitter data respectively were further improved by comprehensive annotation of emotion and sentiment features. And two emotion data sets such as Cleaned Balanced Emotional Tweets (CBET) dataset and Twitter Emotion Corpus (TEC), which is also crawled from Twitter using the Twitter API and labeled automatically using hashtags. Due to the inaccuracy of the hashtag labeling, a procedure was then carried out to validate the annotation of the emotion dataset labels. The validated dataset was then used to train the emotion detection model (EDM) using BERT as a pre-trained word representation model. Subsequently, the extracted emotions were used as features in addition to contextual and sentiment features are fed to deep learning models to train cyberbullying detection models. A set of experiments were carried out with different selections of features to investigate the best set of features for cyberbullying detection. The proposed models is also compared with the state-of-the-art models.

ADVANTAGES

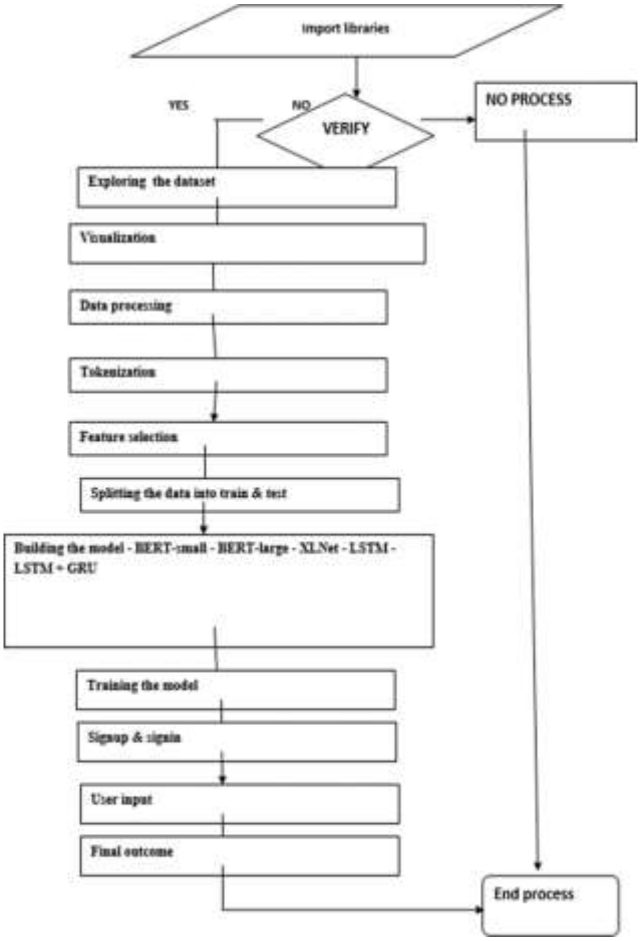
In contrast, our work provides a more comprehensive approach by incorporating contextual, emotion, and sentiment features for cyberbullying detection. This wider range of features can provide a more nuanced understanding of abusive behavior.

We integrates emotion detection using advanced models like BERT, which enables a better identification of emotional content associated with cyberbullying.

Our work indicates that the emotion and sentiment annotations in the datasets were comprehensively improved. This annotation enhancement could potentially lead to more accurate and informative insights into cyberbullying behavior.

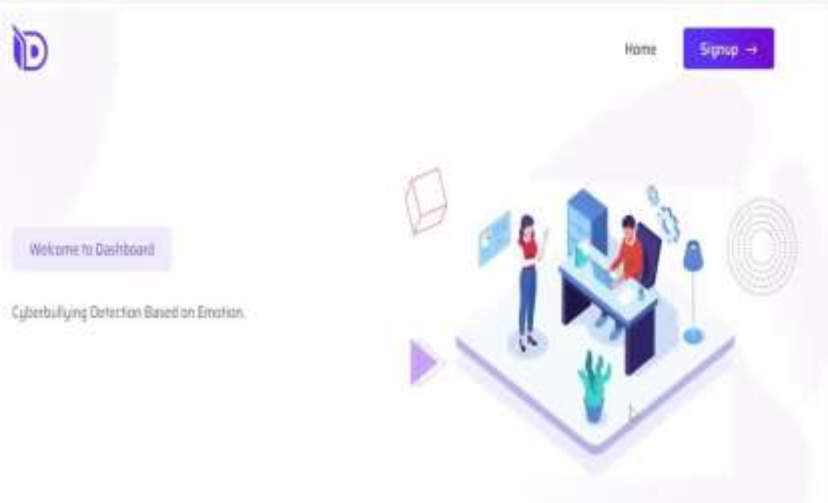
Our work suggests the use of deep learning models trained on a combination of emotional, contextual, and sentiment features. This approach may result in improved accuracy in cyberbullying detection.

System Block Diagram

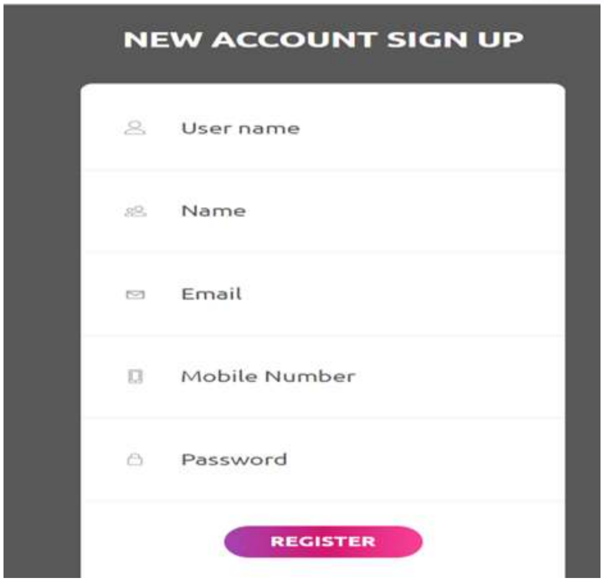


OUTPUTSCREENS

Fig1:Home Page



Cyberbullying Detection Based on Emotion.



NEW ACCOUNT SIGN UP

User name

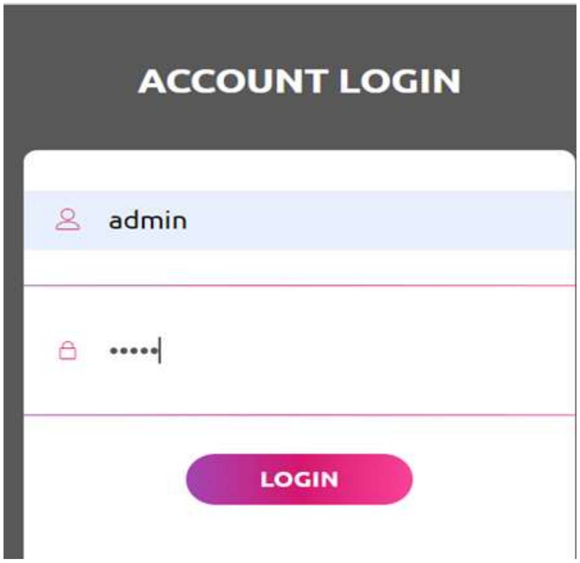
Name

Email

Mobile Number

Password

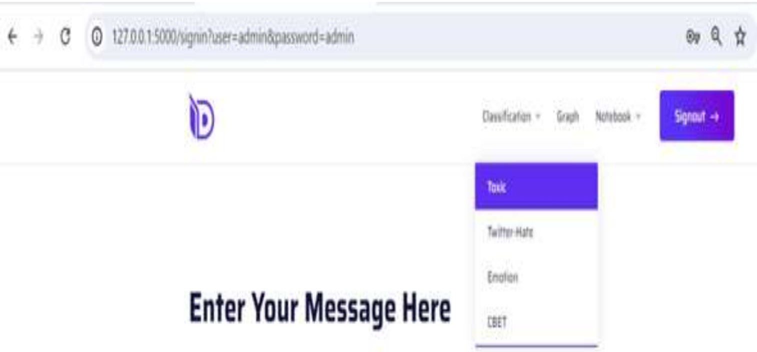
Fig2: SignUp Page



admin

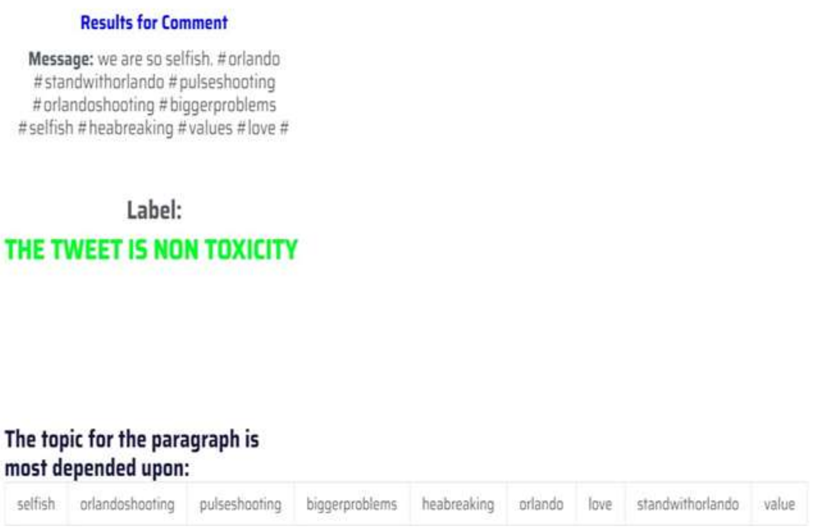
•••••

Fig3:SignIn Page



127.0.0.1:5000/signin?user=admin&password=admin

Fig3:Clasification Page



Results for Comment

Message: we are so selfish. #orlando #standwithorlando #pulseshooting #orlandoshooting #biggerproblems #selfish #heabreaking #values #love #

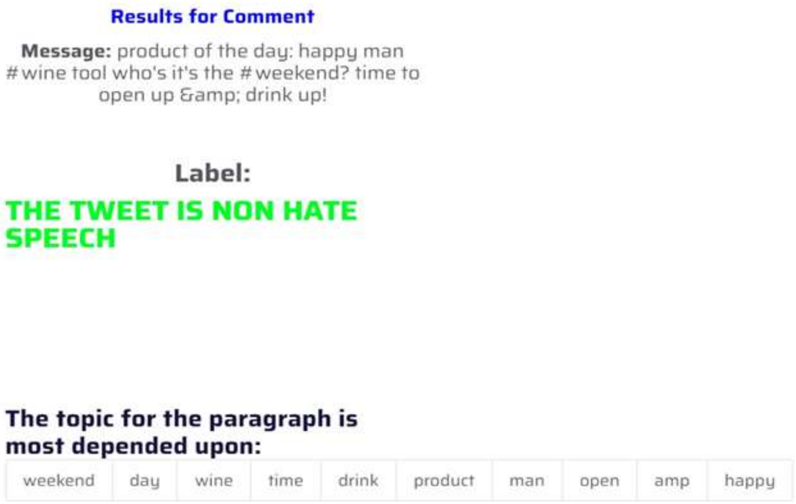
Label:

THE TWEET IS NON TOXICITY

The topic for the paragraph is most depended upon:

selfish	orlandoshooting	pulseshooting	biggerproblems	heabreaking	orlando	love	standwithorlando	value

Fig4:Toxicdata Predicted Page



Results for Comment

Message: product of the day: happy man #wine tool who's it's the #weekend? time to open up &amp; drink up!

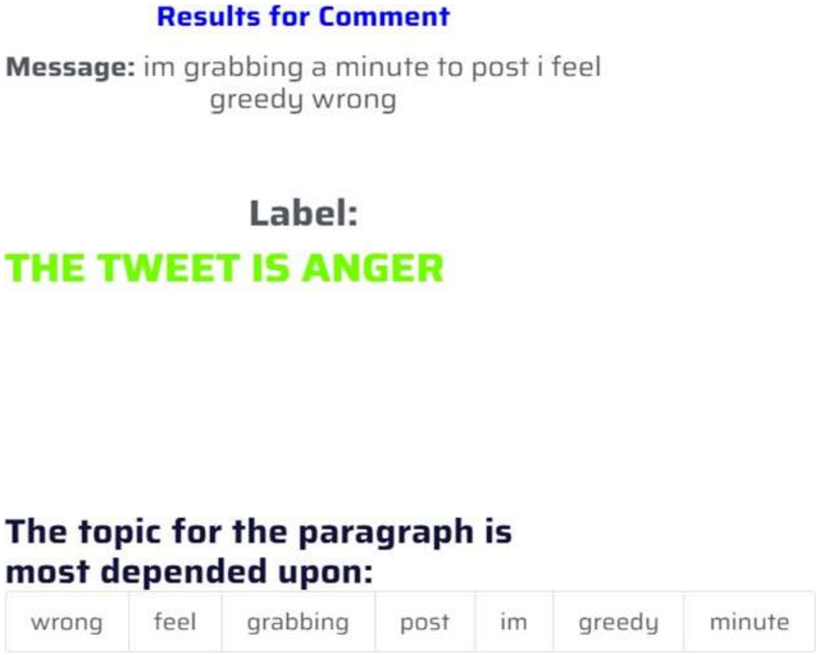
Label:

THE TWEET IS NON HATE SPEECH

The topic for the paragraph is most depended upon:

weekend	day	wine	time	drink	product	man	open	amp	happy

Fig5:Hatespeech Predicted Page



Results for Comment

Message: im grabbing a minute to post i feel greedy wrong

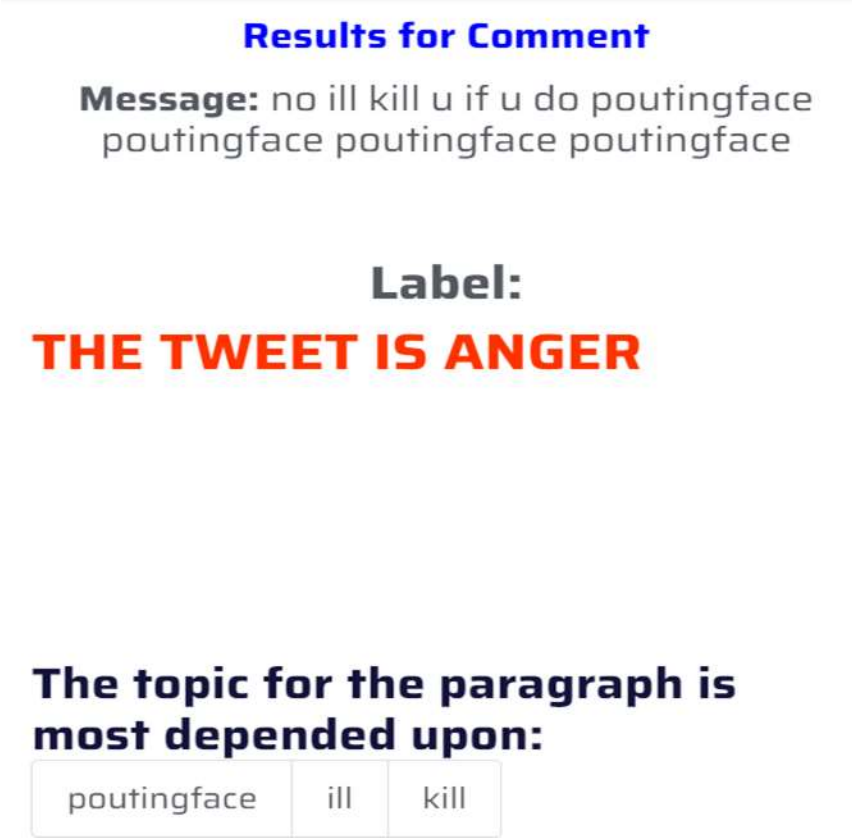
Label:

THE TWEET IS ANGER

The topic for the paragraph is most depended upon:

wrong	feel	grabbing	post	im	greedy	minute

Fig6:Emotion Predicted Page



Results for Comment

Message: no ill kill u if u do poutingface poutingface poutingface poutingface

Label:

THE TWEET IS ANGER

The topic for the paragraph is most depended upon:

poutingface | ill | kill

Fig7:CBET PredictedPage

CONCLUSION

In this work, we proposes a cyberbullying detection models (CDMs) that utilize emotion features to enhance the efficiency of detecting cyberbullying. In this work, all critical steps were taken into consideration, from data preparation to deep learning models.There is a sparsity issue in cyberbullying datasets that encompasses all forms of cyberbullying, such as threatening, harassing, humiliating, intimidating, and manipulating or controlling targeted victims. To address this issue, this research utilized two datasets. The first is the toxic dataset collected by the Conversation AI team, and the second is the Twitter dataset. The limitations of sparsity and imbalance in the cyberbullying dataset were addressed and resolved. After the preparation of the datasets, extracting textual features was the second step in detecting cyberbullying. To build an emotion detection model, the CBET dataset which was collected from Twitter using hashtag keywords was used. The dataset was labeled using hashtags as the keywords. Due to the inaccuracy of the hashtag labeling, a procedure was then carried out to validate the annotation of the emotion dataset labels. The validated dataset was then used to train the emotion detection model (EDM) using BERT as a pre- trained word representation model. This model was used to study and explore the emotions related to cyberbullying texts. The results indicate that most cyberbullying texts are categorized as negative emotions. Emotions and sentiment were drawn out from cyberbullying datasets through the use of EDM and NRC lexicon for emotions and AFINN lexicon for sentiment. These features were fed to deep learning models to train cyberbullying detection models. A set of experiments were carried out with different selections of features to investigate the best set of features for cyberbullying detection. The results show that emotions and sentiment features improve the precision of cyberbullying detection and outperformed the use of BERT contextual features. The use of emotion features added to BERT resulted in a good recall score on the Toxic dataset, enhancing the performance of cyberbullying detection compared to using BERT alone. While the use of sentiment features scored good recall, improving the model compared to using BERT alone which in general is greater than the baseline.

REFERENCES

[1] S. Modha, P. Majumder, T. Mandl, and C. Mandalia, ‘‘Detecting and visualizing hate speech in social media: A cyber watchdog for surveillance,’’ Expert Syst. Appl., vol. 161, Dec. 2020, Art. no. 113725.

[2] S. Hinduja and J. W. Patchin, ‘‘Bullying, cyberbullying, and suicide,’’ Arch. Suicide Res., vol. 14, no. 3, pp. 206–221, Jul. 2010.

[3] R. M. Kowalski, G.W. Giumetti, A.N. Schroeder, M.R. Lattanner, ‘‘Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth,’’ Psychol. Bulletin, vol. 140, p. 1073, 2014, doi: 10.1037/a0035618.

[4] V. Balakrishnan, ‘‘Cyberbullying among young adults in Malaysia: The roles of gender, age and internet frequency,’’ Comput. Hum. Behav., vol. 46, pp. 149–157, May 2015.

[5] S. M. B. Bottino, C. M. C. Bottino, C. G. Regina, A. V. L. Correia, and W. S. Ribeiro, ‘‘Cyberbullying and adolescent mental health: Systematic review,’’ Cadernos Saude Publica, vol. 31, no. 3, pp. 463–475, Mar. 2015