10.48047/jocaaa.2024.32.02.43

SENTIMENTAL CLASSIFICATION-OPTIMIZED TOPIC- ENHANCED NEURAL LANGUAGE MODEL

DR.V.ANANTHA KRISHNA[1], REDDY LAHARI[2], SHIVANI NAGAPURI[2],

THUDIMALLA SHILPA[4]

[1]Professor, Department of Computer Science and Engineering, Sridevi Women's Engineering College, Hyderabad, India

Email: krishnaananthav@gmail.com

[2,3,4.]B.Tech Student, Department of Computer Science and Engineering, Sridevi Women's Engineering College, Hyderabad, India

Abstract: Information and Communication Technologies fueled social networking and facilitated communication. However, cyberbullying on the platform had detrimental ramifications. The user-dependent mechanisms like reporting, blocking, and removing bullying posts online are manual and ineffective. We propose the development of an automatic system for cyberbullying detection using two approaches: Conventional Machine Learning and Transfer Learning. We used the AMiCA dataset, which contains a significant amount of cyberbullying context and structured annotation process. We used various features like textual, sentiment and emotional, static and contextual word embeddings, psycholinguistics, term lists, and toxicity features to detect cyberbullying. The contextual embeddings of ggeluBert, tnBert, and Distil Bert have alike performance, however Distil Bert embeddings were elected for higher F-measure. Textual features, DistilBert embeddings, and toxicity features that struck new benchmark were the top three unique features when fed individually. The model's performance was boosted to F-measure after feeding with a combination of textual, sentiment, Distil Bert embeddings, psycholinguistics, and toxicity features to the Logistic Regression model that outperforms Linear SVC with faster training time and efficient handling of high-dimensionality features. Transfer Learning approach was by fine-tuning optimized version Pre-trained Language Models namely, DistilBert, DistilRoBerta, and Electra-small which were found to have speedier training computation thantheir base form. The fine-tuned DistilBert resulted with the highest F-measure, surpassing CML. Our research concluded that Transfer Learning was the best for uplifted performance and lesser effort as feature engineering and resampling was omitted.

Keywords: Machine Learning, Social Media, Transfer earning, Natural Language Processing, Textual Data, Opinion Mining, Logistic Regression

Index Terms *- Cyberbullying, Automatic cyberbullying detection, Machine Learning, Transfer Learning, AMiCA dataset, Textual features, Sentiment and emotional features, Contextual word embeddings, Toxicity features, DistilBert embeddings.*

1. INTRODUCTION

The 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

10.48047/jocaaa.2024.32.02.43

people easily misuse technological advancement with abusive behaviors such as cyberbullying. Cyberbullying is the expanded form of director traditional bullying through electronic platforms. Social media becomes the virtual medium for bullying, shielding the bully's identity, making detecting cyber bullying a complex and challenging mission to protect online communities. Cyberbullying cases increase with volumized Internet usage because it can be easily committed anonymously, leading to a grave public health concern that brings many negative impacts, such as mental, psychological, and social problems.

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.

Cyberbullying has become a pervasive issue affecting adolescents worldwide, and its impact on mental health and well-being has garnered increasing scholarly attention. Studies such as those by Cagirkan and Bilek [1], Chi et al. [2], and Kowalski and Limber [4] have explored the prevalence and correlates of cyberbullying among high school students. The rise of online platforms and social media has exacerbated the problem, as evidenced by research conducted by López-Martínez et al. [3], Anwar et al. [6], and Garrett et al. [9]. The global pandemic, as investigated by Kee et al. [7], has further amplified cyberbullying on social media platforms, highlighting the need for comprehensive understanding and effective preventive measures.

Addressing cyberbullying requires innovative approaches, as showcased in studies like those of Ptaszynski et al. [11], Raza et al. [12], and Nguyen et al. [13], which delve into the application of machine learning and deep learning models for cyberbullying detection. Moreover, the exploration of psychological theories, as done by Huang [5], contributes to understanding the underlying factors that may contribute to cyberbullying behaviors.

This introduction provides a snapshot of the diverse research landscape on cyberbullying, encompassing its prevalence, psychological correlates, technological interventions, and the evolving challenges posed by the digital age. The subsequent analysis of these studies promises to shed light on the multifaceted

10.48047/jocaaa.2024.32.02.43

nature of cyberbullying and inform strategies for prevention and intervention in both educational and online environments.

2. LITERATURE REVIEW

Cyberbullying, a prevalent issue in today's digitally connected society, has gained considerable attention from researchers exploring its manifestations, impacts, and potential mitigation strategies. This literature survey aims to provide an overview of recent research findings on cyberbullying among high school students, detection methods, and the psychological consequences associated with this online phenomenon. Several studies have delved into the prevalence and characteristics of cyberbullying among high school students. Cagirkan and Bilek [1] conducted a study in Turkey, shedding light on the prevalence and forms of cyberbullying.

Chi et al. [2] focused on the relationship between online time, cyberbullying experiences, and coping strategies among high school students in Hanoi. These studies collectively contribute to a global understanding of the issue, emphasizing the need for cross-cultural insights. Detection methods for cyberbullying have also been a focal point in recent research. Lopez-Martinez et al. [3] introduced CyberDect, a novel approach for cyberbullying detection on Twitter, showcasing the importance of technological solutions in identifying and preventing online harassment.

Ptaszynski et al. [10] and Raza et al. [12] explored the use of machine learning in automatic extraction of harmful sentence patterns for cyberbullying detection, demonstrating the potential of technology in mitigating this pervasive issue. The psychological impact of cyberbullying is a critical aspect explored in

10.48047/jocaaa.2024.32.02.43

the literature. Kowalski and Limber [4] investigated the psychological, physical, and academic correlates of both cyberbullying and traditional bullying, providing a comprehensive understanding of the multifaceted consequences. Workplace cyberbullying was studied by Anwar et al. [6], who explored the mediating effect of silence and emotional exhaustion, contributing to our understanding of the broader implications of cyberbullying beyond the school environment. The influence of external factors, such as the COVID-19 pandemic, on cyberbullying has been explored by Kee et al. [7]. This study highlighted the changing dynamics of cyberbullying on social media platforms under the influence of global events, emphasizing the need for adaptive strategies in addressing online harassment. In addition to individual studies, systematic reviews have been conducted to consolidate knowledge on the association between cyberbullying and mental health.

Kwan et al. [8] presented a systematic map of systematic reviews, providing a comprehensive overview of existing evidence on the impact of cyberbullying on children and young people's mental health. Technological advancements in natural language processing and machine learning have been crucial in developing effective detection models. Ptaszynski et al. [11] and Cai et al. [14] showcased the application of deep learning models in extracting harmful sentence patterns and text classification, underlining the potential for technology to play a vital role in combating cyberbullying. Despite advancements, challenges in dealing with cyberbullying persist.

Twitter's CEO acknowledged shortcomings in addressing abuse on the platform [15], emphasizing the ongoing need for both technological and policy interventions to create safer online spaces. In conclusion, this literature survey provides a comprehensive overview of recent research on cyberbullying, encompassing its prevalence, detection methods, psychological impacts, and the role of technology in addressing this complex issue. The findings underscore the urgency of multidisciplinary efforts involving educators, policymakers, and technologists to create a safer online environment for individuals, especially the younger generation.

3. METHODOLOGY

In previous study they developed model to address the issue of cyberbullying by developing a model that can detect cyberbullying in social media posts and comments. We propose the use of a custom word embedding trained on word2vec, upon which an LSTM-CNN architecture is built and trained. Their model is tested on Twitter posts and comments. In another research they introduced an automatic detection of cyberbullying in social media text by modeling posts written by bullies, victims, and bystanders of online bullying. They collected and fine-grained annotated a cyberbullying corpus for English and Dutch and performed a series of binary classification experiments to determine the feasibility of automatic cyberbullying detection. They used linear support vector machines exploiting a rich feature set and investigated which information sources contribute the most for the task.

Drawbacks:

1. The existing work might suffer from limited feature representation as it primarily relies on custom word embeddings and LSTM-CNN architecture.

10.48047/jocaaa.2024.32.02.43

2. It is impossible to handle the vast volume of data on the Internet within a short time without a computational approach

3. The existing work's usage of Twitter data might limit the generalization of the model to other platforms or contexts.

4. Linear support vector machines might be less capable of capturing complex patterns in the data.

5. They did not experimented with different combinations of features, so the optimal model was not identified which leads to decrease in performance

We propose the development of an automatic system for cyberbullying detection using two approaches: Conventional Machine Learning and Transfer Learning. We used the AMiCA dataset, which contains a significant amount of cyber bullying context and structured annotation process. We used various features like textual, sentiment and emotional, static and contextual word embeddings, psycholinguistics, term lists, and toxicity features to detect cyberbullying. The contextual embeddings of ggeluBert, tnBert, and DistilBert have alike performance, however DistilBert embeddings were elected for higher F-measure. Textual features, DistilBert embeddings, and toxicity features that struck new benchmark were the top three unique features when fed individually.

Benefits:

1. Transfer Learning approach was by fine-tuning optimized version Pre-trained Language Models namely, DistilBert, DistilRoBerta, and Electra-small which were found to have speedier training computation than their base form.

2. Our utilizes a more comprehensive dataset that could lead to improved performance on different social media platforms.

3. By using multiple techniques, the model can leverage the strengths of each approach and combine them to achieve better overall performance in cyberbullying detection.

4. Differente combinations of feature were experimented, which further boost the performance of each model.

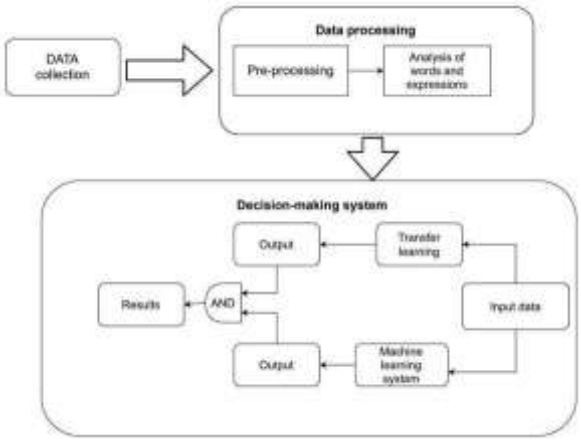

Fig 1 System Architecture

Modules:

The modules are:

- Data exploration: using this module we will load data into system

- Processing: Using the module we will read data for processing

10.48047/jocaaa.2024.32.02.43

- Splitting data into train &test: using this module data will be divided into train & test

- Model generation: Building the model – -ML -LinearSVC -Logistic Regression -Voting CLassifier (AB + RF) -Transfer Learning DistiBert -DistoRoBerta -Electra -DistiBert Embedding -LinearSVC -LogisticRegression -Voting CLassifier -DistiRoBerta -LinearSVC -LogisticRegression -Voting CLassifier -Electra -LinearSVC -LogisticRegression -Voting CLassifier -Deep Learning -LSTM -LSTM + GRU. Algorithms accuracy calculated

- User signup & login: Using this module will get registration and login

- User input: Using this module will give input for prediction

- Prediction: Final predicted displayed

4. IMPLEMENTATION

LinearSVC- Linear Support Vector Machine (Linear SVC) is an algorithm that attempts to find a hyperplane to maximize the distance between classified samples.

Logistic Regression- Logistic regression is a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome, event, or observation. The model delivers a binary or dichotomous outcome limited to two possible outcomes: yes/no, 0/1, or true/false.

Voting CLassifier (AB + RF)- A Voting Classifier is a machine learning model that trains on an ensemble of numerous models and predicts an output (class) based on their highest probability of chosen class as the output.

DistiBert- DistilBERT model is a distilled form of the BERT model. The size of a BERT model was reduced by 40% via knowledge distillation during the pre-training phase while retaining 97% of its language understanding abilities and being 60% faster.

Electra- ELECTRA uses a new pre-training task, called replaced token detection (RTD), that trains a bidirectional model (like a MLM) while learning from all input positions (like a LM). Inspired by generative adversarial networks (GANs), ELECTRA trains the model to distinguish between “real” and “fake” input data.

DistiBert Embedding- DistilBERT is a small, fast, cheap and light Transformer model trained by distilling BERT base. It has 40% less parameters than bert-base-uncased, runs 60% faster while preserving over 95% of BERT's performances as measured on the GLUE language understanding benchmark.

LSTM- It is a special type of Recurrent Neural Network which is capable of handling the vanishing gradient problem faced by RNN. LSTM was designed by Hochreiter and Schmidhuber that resolves the problem caused by traditional rnns and machine learning algorithms. LSTM can be implemented in Python using the Keras library.

LSTM + GRU- the LSTM (Long -short-term memory) and GRU (Gated Recurrent Unit) have gates as an internal mechanism, which control what information to keep and what information to throw out. By doing this LSTM, GRU networks solve the exploding and vanishing gradient problem

10.48047/jocaaa.2024.32.02.43

5. EXPERIMENTAL RESULTS

Dataset Description:

As social media usage becomes increasingly prevalent in every age group, a vast majority of citizens rely on this essential medium for day-to-day communication. Social media's ubiquity means that cyberbullying can effectively impact anyone at any time or anywhere, and the relative anonymity

of the internet makes such personal attacks more difficult to stop than traditional bullying.

On April 15th, 2020, UNICEF issued a warning in response to the increased risk of cyberbullying during the COVID-19 pandemic due to widespread school closures, increased screen time, and decreased face-to-face social interaction. The statistics of cyber bullying are outright alarming: 36.5% of middle and high school students have felt cyber bullied and 87% have observed cyber bullying, with effects ranging from decreased academic performance to depression to suicidal thoughts.

In light of all of this, this dataset contains more than 47000 tweets labeled according to the class of cyber bullying

The data has been balanced in order to contain ~8000 of each class.

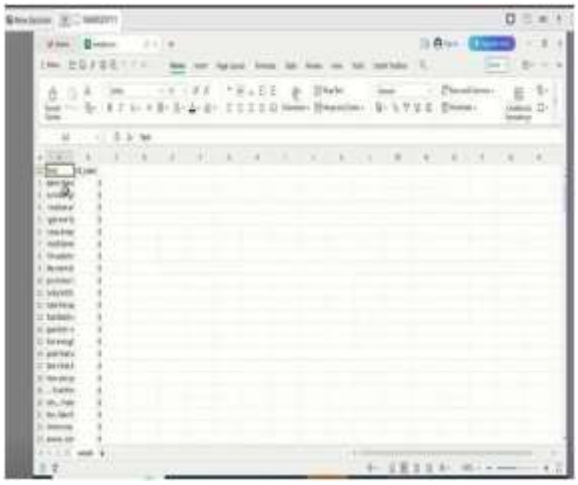

Fig 2 Dataset

Fig 3 Anaconda prompt

Fig 4 URL

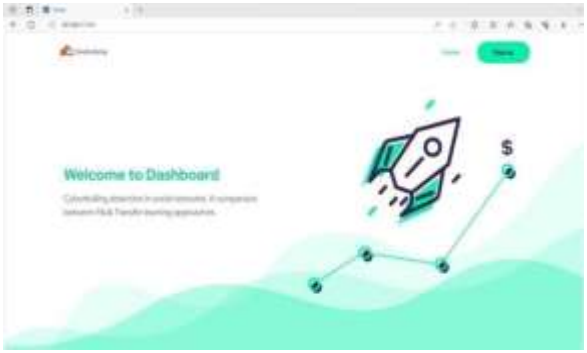

Fig 5 Home page screen

10.48047/jocaaa.2024.32.02.43

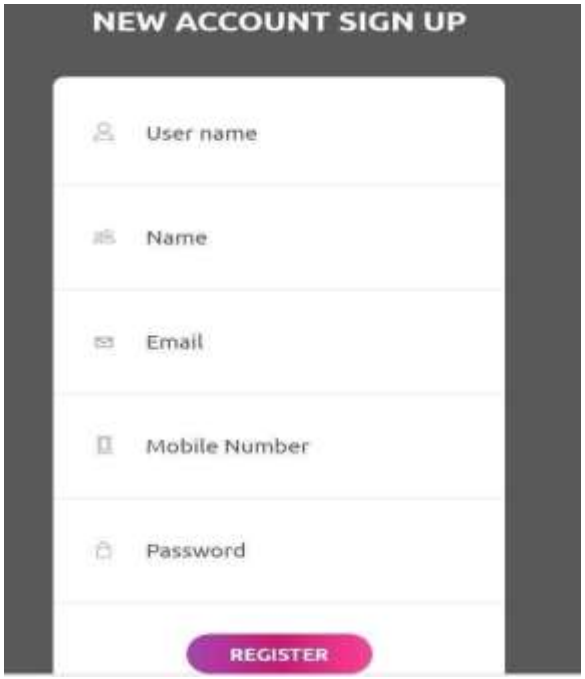

Fig 6 Registration page

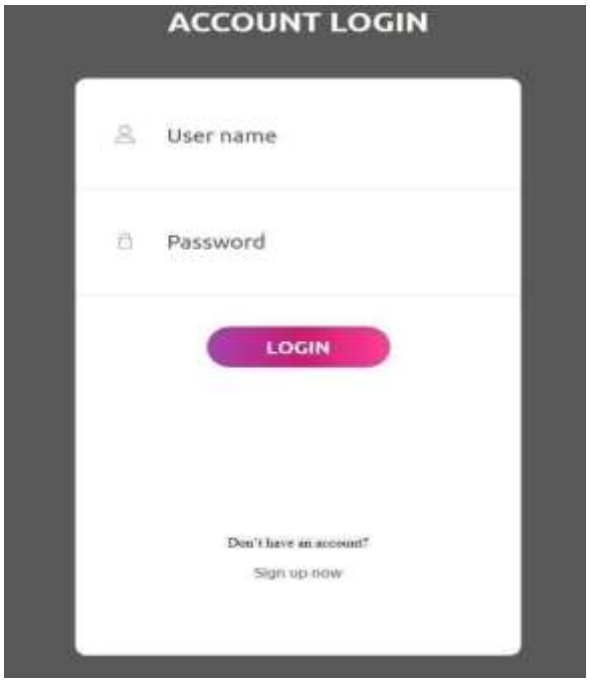

Fig 7 Login page

Fig 8 Input screen

Fig 9 Output screen

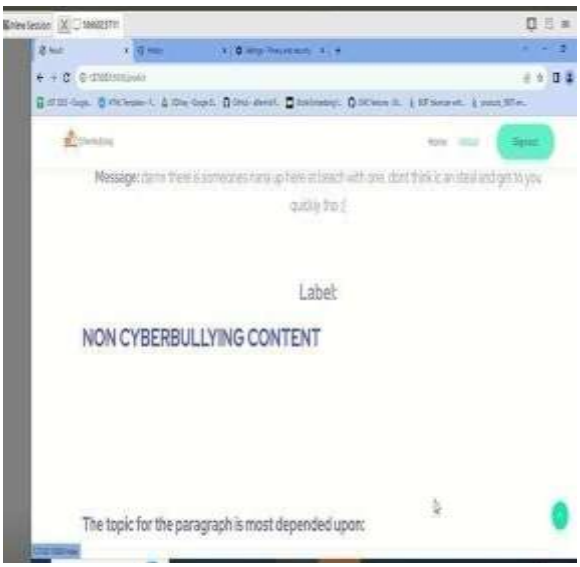

Fig 10 Output screen

10.48047/jocaaa.2024.32.02.43

Fig 11 Output screen

6. CONCLUSION

The conclusion of our work is that cyberbullying is an unexpected ramification of technological advancement, which can bring destructive consequences to any Internet user. Automatic detection is essential for the prevention and reduction to curve out the act from spreading. We explored various features that could be crafted from text and shed light on the methodological steps to adopt textual features, sentiment and emotional features, embeddings, psycholinguistics features, term lists features, and toxicity features. The feature engineering process was part of the conventional machine learning approach. We developed different cyberbullying detection models using Logistic Regression and Linear SVC by conventional machine learning approach and by fine-tuning DistilBert, DistilRoBerta, and Electra-small via epochs training to attain the task, which was a form of binary text classification. We concluded that the proposed feature sets were a step forward for the conventional machine learning approach in working with cyberbullying detection. The result was proven to be optimistic with these features. The highest F-measure was obtained by coupling Logistic Regression with SMOTE resampling and a combination of textual features, sentiment and emotion features, DistilBert embeddings, psycholinguistics features, and toxicity features (A + B + C + D + F) was good for cyberbullying detection. We also supported that toxicity detection should be incorporated into cyberbullying detection from the perspective of model evaluation.

FUTURE WORK:

In terms of future enhancements for comparing transfer learning and machine learning strategies for cyberbullying detection on social networks, there are a few potential areas to explore. One possibility is to investigate the effectiveness of different pre-training models intransfer learning, such as BERT or GPT, and how they can be fine-tuned for cyberbullying detection. Another avenue of research could involve exploring ensemble methods that combine multiple machine learning models to improve the overall performance of cyber bullying detection systems. Additionally, considering the dynamic nature of online content, incorporating real-time data streaming and online learning techniques could be a promising direction for enhancing cyber bullying detection on social networks.

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.

10.48047/jocaaa.2024.32.02.43

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.

10.48047/jocaaa.2024.32.02.43

[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