Journal of Information Systems and Informatics

Vol. 6, No. 2, June 2024 e-ISSN: **2656-4882** p-ISSN: **2656-5935**

DOI: 10.51519/journalisi.v6i2.692 Published By DRPM-UBD

Shielding Social Media: BERT and SVM Unite for Cyberbullying Detection and Classification

Parth Aggarwal[1], Rhea Mahajan[2,*]

[1]Pathways World School Gurgaon, Haryana, India
[2]Department of Computer Science and IT, University of Jammu, J&k, India
Email: [1]aggarwalparth2007@gmail.com, [2]rhea.mahajan@jammuuniversity.ac.in

Abstract

This paper presents a novel approach for cyberbullying detection and classification in social media text using an ensemble model that combines BERT (Bidirectional Encoder Representations from Transformers) and Support Vector Machine (SVM) with grid search for multiclass classification. We have also compared the performance of our proposed with various machine and deep learning models and the results show that our proposed model outperforms other models achieving an accuracy of 90% on testing data. Further, we have used to used SHapley Additive exPlanations (SHAP) an Explainable (XAI) technique to interpret the predictions of the BERT-SVM ensemble model.

Keywords: BERT, SVM, Cyberbully, XAI, SHAP

1. INTRODUCTION

Cyberbullying is a form of harassment that occurs through electronic means, such as online platforms and digital devices. It involves the use of toxic comments, offensive language, hate speeches, or threats targeting an individual's identity. The rise of social media platforms has brought about a concerning increase in cyberbullying incidents. Unfortunately, a significant number of people, particularly teenagers, fall victim to online abuse daily. The consequences of cyberbullying can have a profound and lasting impact on the mental well-being of the targeted individuals. The emotional distress caused by cyberbullying can also result in self-harm behaviors or even lead to depression, a serious mental health condition characterized by persistent feelings of sadness, hopelessness, and a loss of interest in activities. In extreme cases, the relentless harassment and humiliation experienced through cyberbullying can tragically lead to suicide.

In the face of these alarming consequences, there is an urgent imperative to address cyberbullying comprehensively. The creation of a safe and supportive online environment demands not only proactive measures to prevent and mitigate cyberbullying but also educational initiatives to promote empathy, tolerance, and responsible digital citizenship. By fostering a culture of respect and understanding, society can collectively work towards eradicating the scourge of cyberbullying and ensuring the well-being of individuals in the digital age.

Cyberbullying has gained global prevalence and has become an issue of significant concern in countries like India, Brazil, the United States, Belgium and South Africa. India has the highest rate of cyberbullying globally, with more than 85% of children reporting instances of cyberbullying [1]. Furthermore, Indian children reported cyberbullying twice the rate of children worldwide. In India, 46% of teenagers reported being cyberbullied by strangers as compared to the global average of 17%. Additionally, 48% reported being cyberbullied by someone they know, compared to 21% in other countries.[2]. The types of cyberbullying prevalent in India were identified as spreading false rumors (39% of reported cases), being excluded from chats or groups (35%), and name-calling (34%). Over the past decade, researchers have recognized the utmost significance of studying and addressing cyberbullying.

It is also reported that 77% of these victims of cyberbullying are harassed on social media platforms [3]. A substantial number of individuals, particularly adolescents, find themselves subjected to online abuse on a daily basis. The rise of social media platforms, with over 3 billion active users, has significantly transformed communication and information dissemination. These platforms, including Facebook, Twitter, Instagram, and Tinder, facilitate the sharing of multimedia messages among users. During events like the COVID-19 pandemic, social media played a crucial role in disseminating real-time information. Despite the presence of safety centers on some platforms to monitor and control cyberbullying, the issue persists, highlighting the need for more definitive solutions.

Indeed, social media sites can leverage artificial intelligence (AI) and machine learning technologies to proactively detect and remove cyberbullying content before it spreads. By employing AI algorithms and machine learning models, these platforms can analyze the content posted by users and identify patterns and behaviors associated with cyberbullying. Once cyberbullying content is detected, social media platforms can take appropriate actions, such as removing the content, issuing warnings, or implementing temporary or permanent suspensions for users engaging in cyberbullying.

To tackle this challenge, researchers have explored both traditional machine learning models and deep learning models with diverse word embedding techniques. These investigations consistently demonstrated that deep learning models outperformed their traditional counterparts. In this research, a novel approach to cyberbullying detection has been proposed by combining BERT Bidirectional Encoder Representations from Transformers (BERT) and Support Vector Machine (SVM) model aiming to surpass the performance of previous models. The focus is on identifying five primary categories of cyberbullying and comparing the performance of the proposed model with the previous baseline in detecting and classifying such text.

Combining Support Vector Machines (SVM) and BERT models in an ensemble for cyberbullying detection has proven to be a synergistic approach that capitalizes on their complementary strengths. While BERT excels in capturing contextual nuances and semantic understanding, SVM provides interpretability, efficient handling of imbalanced data, and computational resource efficiency. The ensemble leverages SVM's robustness and interpretability alongside BERT's deep learning capabilities, resulting in a model that benefits from the nuanced feature extraction of BERT and maintains SVM's interpretative clarity.

1.1 Background

In recent years, an abundance of research has been dedicated to the critical area of cyberbullying detection and classification. This section aims to provide a comprehensive overview of studies conducted within the last five years, shedding light on advancements and emerging trends in this rapidly evolving field. While recent studies have shifted their paradigm from machine learning algorithms to deep learning algorithms, many authors still consider that machine learning algorithms such as SVM, Adaboost, J48 perform better on text data [4][5][6].

Meanwhile, researchers have emphasised the intricate relationship between personality traits and online behaviour. Balakrishna et al. [7] demonstrated the potential of incorporating Big Five personality traits into a Random Forest classifier for cyberbullying detection on Twitter, achieving remarkable accuracy rates and offering insights for intervention strategies. Dalvi et al. [8] proposed a supervised machine learning-based method to detect and prevent Internet exploitation on Twitter, achieving higher accuracy with Support Vector Machine (SVM)compared to Naive Bayes. Bozyiğit et al. [9] adopted a multifaceted approach to cyberbullying detection on Twitter, incorporating various social media features beyond textual content, and evaluating different classifiers. Finally, Saichandana et al. [10] introduced a novel approach to cyberbullying detection by analyzing audio features in voice tweets, with SVM emerging as the most effective classifier in their study.

In recent years, significant advancements have been made in cyberbullying detection on social media platforms using deep learning. Al-Ajlan et al. [11]

introduced OCDD, a deep learning approach utilizing word vectors for Twitter cyberbullying detection, achieving an accuracy of 82.2%. Yadav et al. [12] and Desai et al. [13] demonstrated the effectiveness of BERT models, with Yadav et al. achieving 98% accuracy on Form spring data and Desai et al. achieving 91.90% accuracy on Twitter data. In 2022, Alabdulwahab et al. [14] compared machine learning and deep learning techniques, with the deep learning model achieving the highest accuracy of 0.96.

Roy et al. [15] explored deep transfer learning for image-based cyberbullying detection, achieving 89% accuracy. Raj et al. [16] proposed a CNN-BiLSTM (Convolutional Neural Network-Bidirectional Long Short-Term Memory) model for multilingual cyberbullying detection on Twitter, achieving accuracies of 92.3% for English, 90.2% for Hindi, and 88.7% for Hinglish. In 2023, Fati et al. [17] combined deep learning, attention mechanisms, and CBOW for robust cyberbullying detection, achieving an accuracy of 0.8649. Paruchuri and Rajesh [18] introduced CyberNet, a hybrid deep CNN model, achieving 97% accuracy. Overall, these studies highlight the continuous efforts and advancements in cyberbullying detection, showcasing the integration of advanced technologies and innovative methodologies to address the complex challenges posed by online harassment.

Building upon these foundational studies, Hassanien el al. [19], Kajić et al. [20] and Hassan et.al [21] Yi& Zubiaga [22] conducted comprehensive reviews and systematic analyses of cyberbullying detection techniques, covering a wide range of methodologies and approaches. These studies provided valuable insights into the state-of-the-art in cyberbullying detection, highlighting the significance of machine learning, deep learning, and sentiment analysis in addressing the complex challenges of online harassment.

1.3 Research gaps

The field of cyberbullying detection is continually evolving. While significant progress has been made, research gaps and challenges still need to be addressed. Diverse datasets are crucial for training models that can generalise well across different languages, cultural contexts, and types of cyberbullying instances. Collecting and curating datasets that represent a broad range of online interactions can help build more robust and inclusive models. Moving beyond binary classification (bullying vs. non-bullying) and adopting multiclass classification can provide a better understanding of cyberbullying. Differentiating between various forms of cyberbullying, such as harassment, impersonation, or hate speech, can improve the precision and relevance of detection models.

In the proposed research, an effort has been made to address the aforementioned gaps by taking a multilabel dataset on which no work has been done previously. We have also proposed a novel model for the detection of cyberbullying tweets. Further, we have also used Explainable AI(XAI) techniques such as SHAP to our model to make it more interpretable by providing insights into how the model arrives at its predictions.

2. METHODS

An ensemble model combining the SVM model and the BERT model has been proposed for the intended research. Ensemble models have been proven to produce state-of-the-art results for various classification problems better as the advantages brought by each model are taken into account when producing an output. Figure 1 shows architecture of the proposed research

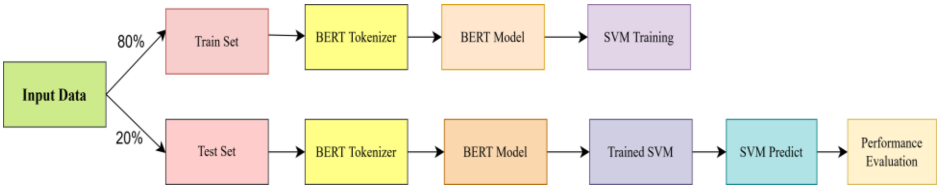
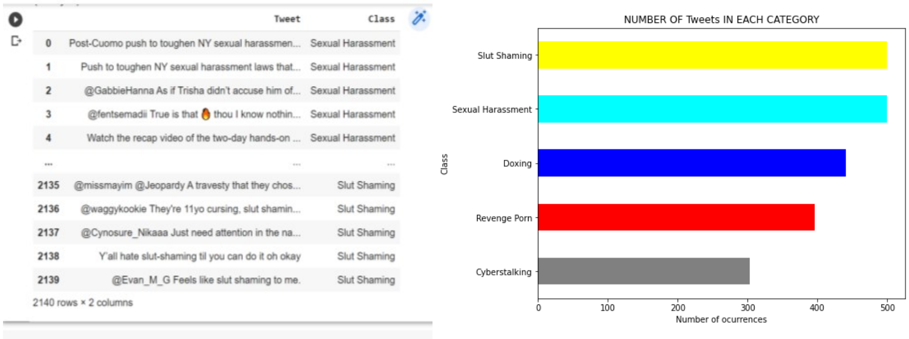


Figure 1. The architecture of the research

BERT, which stands for Bidirectional Encoder Representations from Transformers, is a pre-trained natural language processing (NLP) model. BERT is based on the transformer architecture introduced by Vaswani et al. [26]. Transformers are designed to capture long-range dependencies in sequences efficiently through self-attention mechanisms.



	Tweet	Class
0	Post-Cuomo push to toughen NY sexual harassmen...	Sexual Harassment
1	Push to toughen NY sexual harassment laws that...	Sexual Harassment
2	@GabbieHanna As if Trisha didn't accuse him of...	Sexual Harassment
3	@fentsemadii True is that 🔥 thou I know nothin...	Sexual Harassment
4	Watch the recap video of the two-day hands-on ...	Sexual Harassment
...	...	...
2135	@missmayim @Jeopardy A travesty that they chos...	Slut Shaming
2136	@waggykookie They're 11yo cursing, slut shamin...	Slut Shaming
2137	@Cynosure_Nikaaa Just need attention in the na...	Slut Shaming
2138	Y'all hate slut-shaming til you can do it oh okay	Slut Shaming
2139	@Evan_M_G Feels like slut shaming to me.	Slut Shaming

2140 rows × 2 columns

Figure 2. An instance of data set and number of Tweets in each category

BERT is bidirectional. It considers both the left and right context of each word in a sequence, enabling it to understand the relationships and meanings within a sentence more comprehensively. Whereas, SVM has proven to be a powerful and versatile algorithm with a solid theoretical foundation. Its ability to handle high-dimensional data, support for non-linear decision boundaries through kernels, and effectiveness in various applications make it a popular choice in machine learning.

2.1 Dataset Description

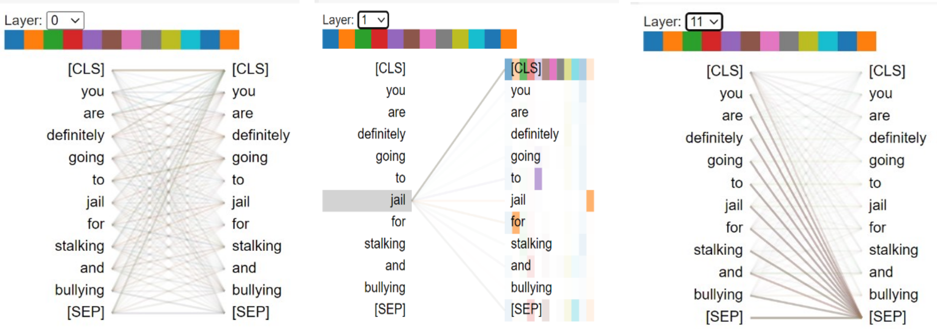
The data set used for the research has been taken from the IEEE data port [23]. This dataset is available in the English language. It is a balanced and labelled data set. It contains 2140 tweets categorized into five categories namely Sexual Harassment, Doxing, Cyberstalking, Revenge porn and Slut Shaming. The number of tweets in each category is shown in Fig 2. After gathering the data, it's essential to undergo a cleaning and processing. The preprocessing step holds significant importance as it directly impacts the effectiveness of subsequent procedures. Our model's preprocessing phase incorporates conventional machine learning preprocessing techniques, such as stop word removal, sentence tokenization, and the removal of special symbols from tweets (like usernames, links, and hashtag symbols). Since we have compared our proposed model with various machine learning models, so Term Frequency-Inverse Document Frequency [24] technique has been for feature engineering then chi-square technique [25] has been used to find most correlated unigrams and bigrams features. These features are listed in Table 1.

Tabel 1. Most correlated Unigrams and Bigrams

Class	Unigrams	Bigrams
Doxing	doxing, private, publish	doxing people, publishing private
Slut shaming	whore, hooker	slut girl, public hooker
Revenge Porn	mms, adult, porno	nude movie, viral https
Sexual Harassment	sexy, movie, sexual	sexy bitch, sexual harassment
Cyberstalking	Stalking, follow, criminal	scary criminal, online stalking

2.2 Methodology

For the research, we have split the dataset into training and testing sets with an 80-20 ratio using scikit-learn's train_test_split function. BERT tokenizer and model are loaded from the 'bert-base-uncased' variant. The Bert tokenizer tokenizes input text into subword tokens using WordPiece tokenization. This involves breaking down words into smaller units and representing them as tokens. Special tokens like [CLS] (classification), [SEP] (separator), and [MASK] (mask) are added to mark the beginning and end of sentences, separate sentences, and handle masked language model tasks, respectively. The Bert tokenizer encodes the texts into input IDs and attention masks, necessary for feeding into the BERT model. Then the BERT model has been used to obtain embeddings (vector representations) for the tokenized texts. These embeddings are then averaged across tokens to get a single vector representation for each text. BERT base model generates high-dimensional vector representations (embeddings) for each input token in a sentence. These embeddings capture rich semantic information about the words in the context of the entire sentence. They are typically used as features.



(a)
(layer 0, head 0)

(b)
(layer 1, head 1)

(c)
(layer 11, head 11)

Figure 3. The attention-head view

The BERT model inputs consist of token embeddings, segment embeddings, and position embeddings. For each input token Xi, a word embedding vector Ei is looked up from an embedding table. BERT uses segment embeddings to distinguish between different sentences in a sequence. Each token is assigned a segment embedding based on the sentence it belongs to. Positional encodings are added to the token embeddings to convey the position information of each token in the sequence. The input representation for each token is the sum of its token embedding, segment embedding, and position embedding. BERT is structured with multiple tiers of Transformers, each layer featuring self-attention mechanisms and feed-forward neural networks. Through self-attention, BERT evaluates the significance of each word or token within the complete input sequence, establishing attention scores between all token pairs and amalgamating information from pertinent tokens. We utilized BertViz, an accessible tool for illustrating multi-head self-attention within the BERT model. BertViz encompasses three perspectives: an attention-head view, a model view, and a neuron view, exemplified by a randomly selected tweet from the dataset [27].

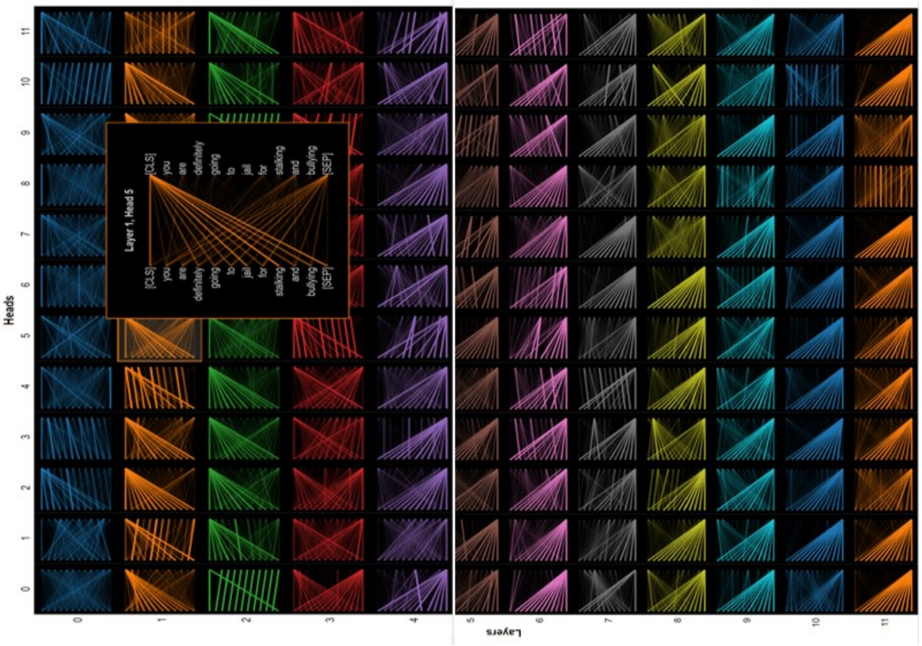


Figure 4. The model view

As depicted in Figure 4, the attention-head view illustrates attention patterns generated by one or more attention heads within a specific transformer layer. This view helps understand which parts of the input text receive more attention during the model's processing, The model view provides a broader perspective by offering a comprehensive overview of attention across all layers and heads of the BERT model. This view enables us to analyse the flow of attention throughout the entire model architecture, as shown in Figure 5. The neuron view, depicted in Figure 5, visualises individual neurons in the query and key vectors of the model. By elucidating the interactions between these neurons and their contributions to producing attention scores, this view helps researchers understand how specific components within the model influence the attention mechanism.

The output of the self-attention mechanism is passed through feed-forward neural networks, which apply non-linear transformations to the input. BERT embeddings are then obtained for the test set and fed into the SVM Classifier which has been instantiated and trained on the training embeddings. We have also applied grid search on an SVC (Support Vector Classifier) to explore all specified combinations of hyperparameters, ensuring that the best hyperparameters within the defined search space are not missed. Finally, the performance of the model has been evaluated using metrics such as precision, recall, and F1-score by plotting a confusion matrix.

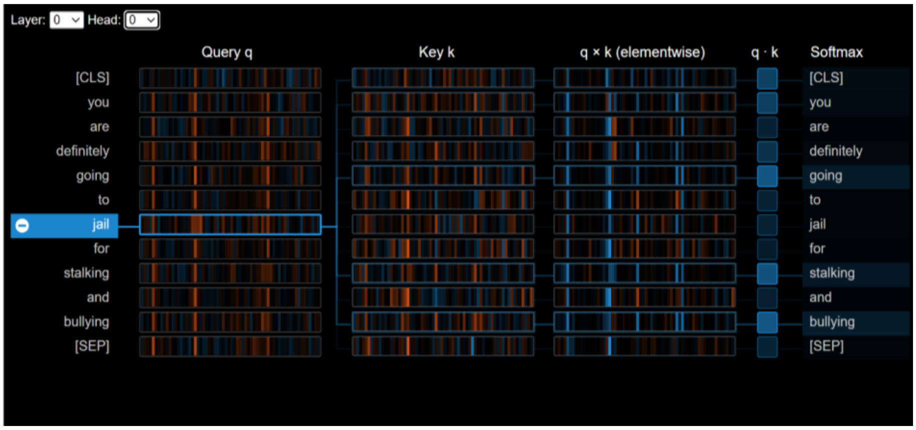


Figure 5. The neuron view

The BERT (Bidirectional Encoder Representations from Transformers) base model is a pre-trained model developed by Google AI's research team. It belongs to the transformer-based language model family and is specifically designed for natural language processing (NLP) tasks. BERT base model consists of a multi-layer bidirectional Transformer encoder architecture. It comprises multiple Transformer blocks stacked on top of each other. Each Transformer block consists of a self-attention mechanism and a feed-forward neural network. These blocks allow BERT to capture the contextual relationships between words in a sentence efficiently.

BERT base model is pre-trained on large corpora of text data using two unsupervised learning tasks: Masked Language Model (MLM) where BERT randomly masks some of the input tokens and tries to predict them based on the surrounding context. This task encourages the model to learn bidirectional representations of the text and Next Sentence Prediction (NSP) where BERT is trained to predict whether two input sentences are consecutive or not. This task helps the model understand the relationships between sentences and improves its ability to capture discourse-level information. Pre-training on these tasks enables BERT to learn rich, contextualized representations of words and sentences, capturing various linguistic patterns and semantic relationships. After pre-training, the BERT base model can be fine-tuned on downstream NLP tasks such as text classification, named entity recognition, question answering, etc. The pre-trained BERT model is further trained on task-specific labelled data during fine-tuning. Fine-tuning adjusts the model's parameters to suit the specific task better, leveraging the contextualised representations learned during pre-training.

3. RESULTS AND DISCUSSION

We have also compared our model with other machine learning models such as Random Forest, SVM, Naïve Bayes and Logistic Regression and deep learning models such as BiLSTM and BERT and comparative analysis show that our model outperformed other baseline models achieving an accuracy of 90% on testing data. Table 2 shows the comparison of our model with other baseline models on various metrics such as precision recall and F1-score along with accuracy. Figure 6 shows the confusion matrix of our proposed ensemble model. The confusion matrix of the models against which our model has been compared is shown in Figure 7.

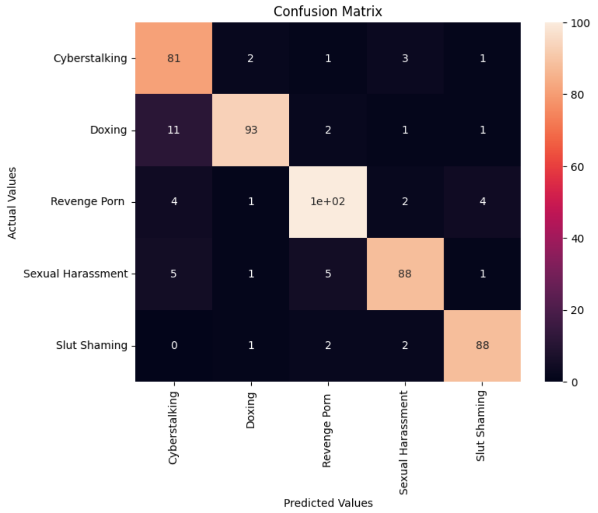
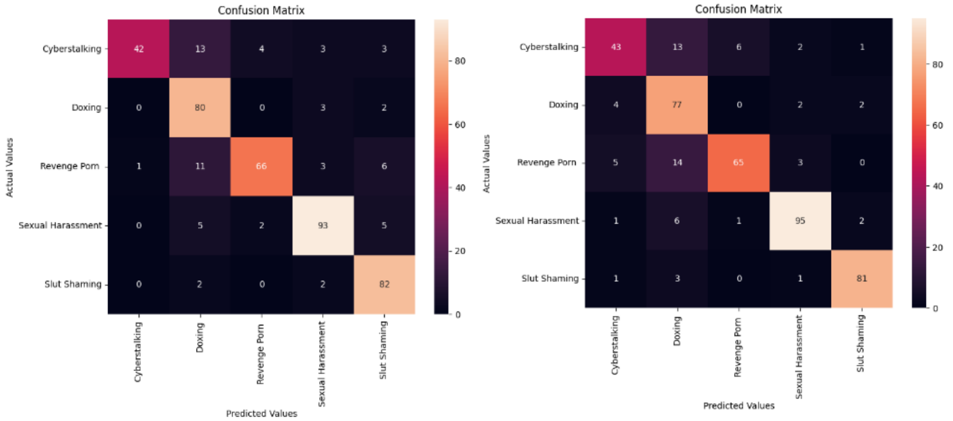
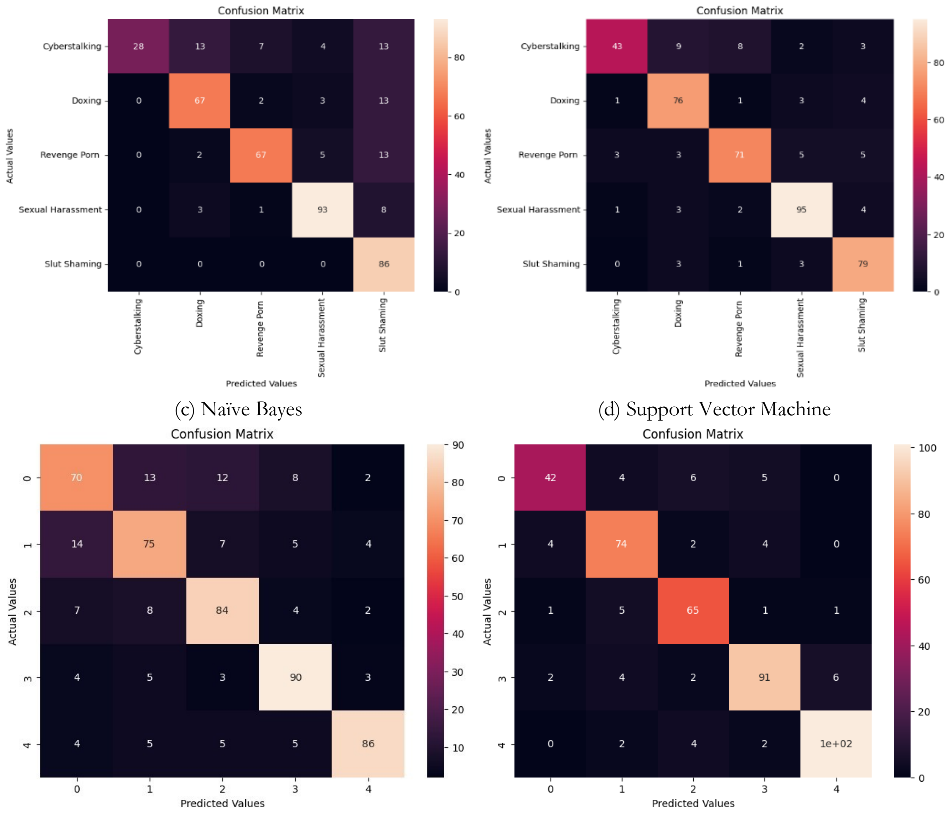


Figure 6. Confusion Matrix of the proposed model (BERT+SVM)



(a) Logistic Regression

(b) Random Forest



(c) Naïve Bayes

(d) Support Vector Machine

(e) BiLSTM

(f) BERT

Figure 7. Performance Evaluation

Table 2. Comparison of our model with other baseline models

Classifier	Class	Precision	Recall	F1 Score	Model Accuracy
Random Forest	**Cyberstalking**	0.80	0.66	0.72	0.84
	Doxing	0.68	0.91	0.78	
	Revenge Porn	0.90	0.75	0.82	
	Sexual Harrasment	0.92	0.90	0.91	
	Slut Shaming	0.94	0.94	0.94	
Linear SVC	**Cyberstalking**	0.90	0.66	0.76	0.85
	Doxing	0.81	0.89	0.85	
	Revenge Porn	0.86	0.82	0.84	
	Sexual Harrasment	0.88	0.90	0.89	
	Slut Shaming	0.83	0.92	0.87	

Multinomial Naïve Bayes	**Cyberstalking**	1.00	0.43	0.60	0.80
	Doxing	0.79	0.79	0.79	
	Revenge Porn	0.87	0.77	0.82	
	Sexual Harrasment	0.89	0.89	0.89	
	Slut Shaming	0.65	1.00	0.79	
Logistic Regression	**Cyberstalking**	0.98	0.65	0.78	0.85
	Doxing	0.72	0.94	0.82	
	Revenge Porn	0.92	0.76	0.83	
	Sexual Harrasment	0.89	0.89	0.89	
	Slut Shaming	0.84	0.95	0.89	
BLSTM (Bidirectional long Short-Term Memory)	**Cyberstalking**	0.73	0.87	0.79	0.85
	Doxing	0.87	0.85	0.86	
	Revenge Porn	0.82	0.80	0.81	
	Sexual Harrasment	0.89	0.92	0.91	
	Slut Shaming	0.94	0.80	0.87	
BERT	**Cyberstalking**	0.86	0.74	0.79	0.87
	Doxing	0.83	0.88	0.86	
	Revenge Porn	0.82	0.89	0.86	
	Sexual Harassment	0.88	0.87	0.88	
	Slut Shaming	0.94	0.93	0.93	
BERT+SVM (Proposed Method)	**Cyberstalking**	0.80	0.92	0.86	0.90
	Doxing	0.95	0.86	0.90	
	Revenge Porn	0.91	0.90	0.90	
	Sexual Harrasment	0.92	0.88	0.90	
	Slut Shaming	0.93	0.95	0.94	

3.1 SHAP (SHapley Additive exPlanations)

SHAP, stands for "SHapley Additive exPlanations," is a formidable tool within the sphere of explainable AI (XAI). It operates as a model-agnostic approach, providing invaluable insights into the inner workings of predictive models by attributing values to individual features in a prediction. Through this method, SHAP illuminates the contributions of each feature to the model's overall output, effectively unveiling the underlying rationale behind its predictions [28].

In this section, we have explained predictions made by our model by generating SHAP values using Partition Explainer available from SHAP. The Partition Explainer operates by iteratively exploring various hierarchical arrangements of the data's features to determine SHAP values. This process involves recursively dissecting the dataset across different feature dimensions, enabling a comprehensive understanding of how each feature contributes to the model's predictions.

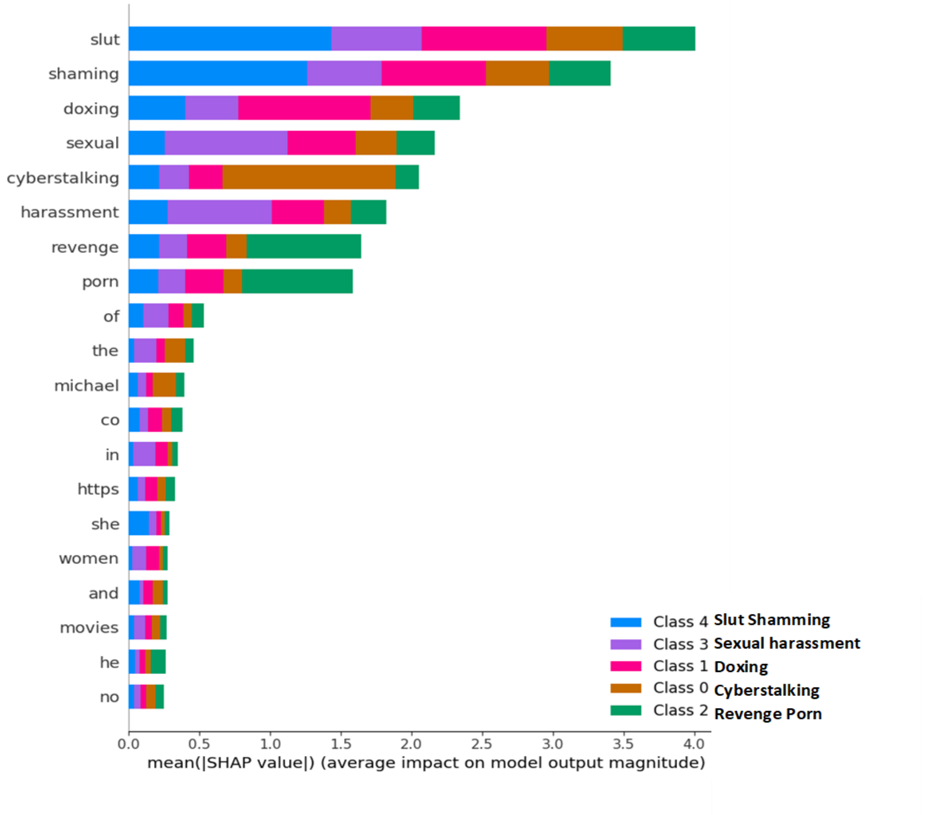
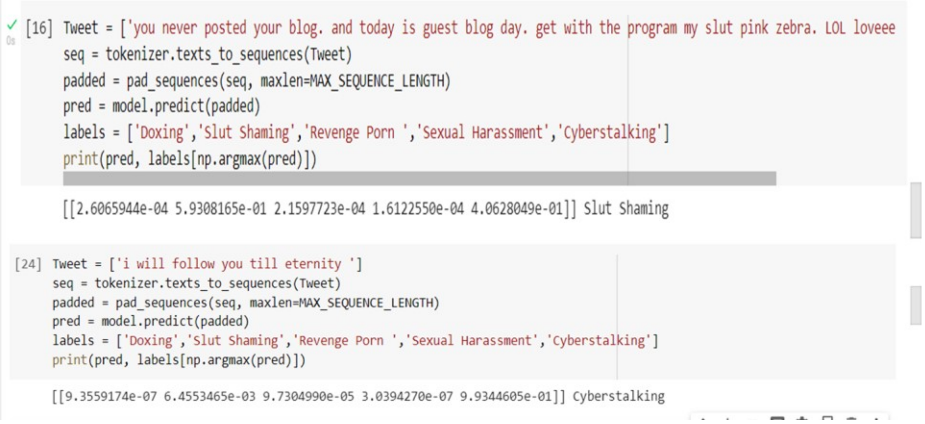


Figure 8. Visual representation of the impact of each feature on a particular prediction made by the proposed model

The bar plot of SHAP (SHapley Additive exPlanations) values is in Figure 8. provide a visual representation of the impact of each feature on a particular prediction made by a machine learning model. Each bar in the plot corresponds to a specific feature, and the length or height of the bar indicates the magnitude and direction of that feature's contribution to the model's prediction. By comparing the heights of different bars, we can discern the relative importance of various features (slut, shaming, doxing etc.,) in shaping the model's prediction as features with taller bars exert a more substantial influence on the prediction compared to those with shorter bars. Figure 9 shows the prediction results of the proposed model.



```
[16] Tweet = ['you never posted your blog. and today is guest blog day. get with the program my slut pink zebra. LOL loveee
     seq = tokenizer.texts_to_sequences(Tweet)
     padded = pad_sequences(seq, maxlen=MAX_SEQUENCE_LENGTH)
     pred = model.predict(padded)
     labels = ['Doxing','Slut Shaming','Revenge Porn ','Sexual Harassment','Cyberstalking']
     print(pred, labels[np.argmax(pred)])

     [[2.6065944e-04 5.9308165e-01 2.1597723e-04 1.6122550e-04 4.0628049e-01]] Slut Shaming
```

```
[24] Tweet = ['i will follow you till eternity ']
     seq = tokenizer.texts_to_sequences(Tweet)
     padded = pad_sequences(seq, maxlen=MAX_SEQUENCE_LENGTH)
     pred = model.predict(padded)
     labels = ['Doxing','Slut Shaming','Revenge Porn ','Sexual Harassment','Cyberstalking']
     print(pred, labels[np.argmax(pred)])

     [[9.3559174e-07 6.4553465e-03 9.7304990e-05 3.0394270e-07 9.9344605e-01]] Cyberstalking
```

Figure 9. Prediction Results

4. CONCLUSION

In this paper, we have introduced a pioneering methodology for the detection and classification of cyberbullying in social media text. Our approach employs an ensemble model that combines the power of BERT (Bidirectional Encoder Representations from Transformers) and Support Vector Machine (SVM), augmented with grid search for multiclass classification. Through comprehensive experimentation and evaluation, we demonstrated the effectiveness of our proposed ensemble model, showcasing its ability to achieve an impressive accuracy rate of 90% on testing data by leveraging the strengths of both deep learning and traditional machine learning techniques.

Moreover, to enhance the interpretability of our model's predictions, we integrated SHAP (SHapley Additive exPlanations), an Explainable AI (XAI) technique, into our framework. By leveraging SHAP, we gained valuable insights into the underlying rationale behind the predictions generated by the BERT-SVM ensemble model, thus facilitating a deeper understanding of the factors influencing cyberbullying classification outcomes. Though our model performs well as compared to other models but it has high computational complexity due to the computational demands of BERT, SVM with grid search, ensemble model construction, and SHAP analysis.

In addition to its academic significance, the study holds significant social and practical implications. The insights gained from the research shed light on the critical need to address the pervasive issue of cyberbullying in contemporary society. By highlighting effective methods for detecting and combating cyberbullying, the findings have the potential to capture the attention of lawmakers and policymakers, prompting action towards implementing necessary measures for prevention and intervention. By safeguarding potential victims from the harmful effects of cyberbullying, such as psychological distress and social isolation, these efforts can contribute to fostering a healthier and more resilient society. Looking ahead, future work in this area aims to expand the scope of analysis by incorporating image data, thus offering a more comprehensive understanding of online behavior and enabling more robust strategies for addressing cyberbullying across various digital platforms. This holistic approach underscores the importance of multidisciplinary research and collaborative efforts in tackling complex societal challenges in the digital age.

REFERENCES

[1] D. Mukhopadhyay, K. Mishra, K. Mishra, and L. B. Tiwari, "Cyber bullying detection based on Twitter dataset," *in Lecture notes in networks and systems (Online)*, 2020, pp. 87–94. doi: 10.1007/978-981-15-7106-0_9.

[2] P. Sangani, "Cyberbullying in children more widespread in India than elsewhere," The Economic Times, Sep. 04, 2022.

[3] O. Djuraskovic, "Cyberbullying Statistics, Facts, and Trends (2023) with Charts," *First Site Guide*, Jun. 2023.

[4] S. Neelakandan et al., "Deep learning approaches for cyberbullying detection and classification on social media," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–13, Jun. 2022, doi: 10.1155/2022/2163458.

[5] A. Perera and P. Fernando, "Accurate cyberbullying detection and prevention on social media," *Procedia Computer Science*, vol. 181, pp. 605–611, Jan. 2021, doi: 10.1016/j.procs.2021.01.207.

[6] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," *Int. Conf. Distrib. Comput. Netw.*, Jan. 2016, doi: 10.1145/2833312.2849567.

[7] V. Balakrishnan, S. Khan, T. Fernandez, and H. R. Arabnia, "Cyberbullying detection on twitter using Big Five and Dark Triad features," *Personality and Individual Differences*, vol. 141, pp. 252–257, Apr. 2019, doi: 10.1016/j.paid.2019.01.024.

[8] R. R. Dalvi, S. Baliram Chavan and A. Halbe, "Detecting A Twitter Cyberbullying Using Machine Learning," *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, 2020, pp. 297-301, doi: 10.1109/ICICCS48265.2020.9120893.

[9] A. Bozyiğit, S. Utku, and E. Nasıbov, "Cyberbullying detection: Utilizing social media features," *Expert Systems with Applications*, vol. 179, p. 115001, Oct. 2021, doi: 10.1016/j.eswa.2021.115001.

[10] B. Saichandana and P. Kamakshi, "Classification of Cyberbullying Detection in Social Networking with Audio using Machine Learning Approach," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, no. 7s, pp. 423–429, Jul. 2023, doi: 10.17762/ijritcc.v11i7s.7018.

[11] M. A. Al-Ajlan and M. Ykhlef, "Optimized Twitter Cyberbullying Detection based on Deep Learning," *2018 21st Saudi Computer Society National Computer Conference (NCC)*, Riyadh, Saudi Arabia, 2018, pp. 1-5, doi: 10.1109/NCG.2018.8593146.

[12] J. Yadav, D. Kumar, and D. Chauhan, "Cyberbullying Detection using Pre-Trained BERT Model," *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, Jul. 2020, doi: 10.1109/icesc48915.2020.9155700.

[13] A. Desai, S. Kalaskar, O. Kumbhar, and R. Dhumal, "Cyber Bullying Detection on Social Media using Machine Learning," *ITM Web of Conferences*, vol. 40, p. 03038, Jan. 2021, doi: 10.1051/itmconf/20214003038.

[14] A. Alabdulwahab, M. A. Haq, and M. S. Alshehri, "Cyberbullying Detection using Machine Learning and Deep Learning," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 10, Jan. 2023, doi: 10.14569/ijacsa.2023.0141045.

[15] P. K. Roy and F. U. Mali, "Cyberbullying detection using deep transfer learning," *Complex & Intelligent Systems*, vol. 8, no. 6, pp. 5449–5467, May 2022, doi: 10.1007/s40747-022-00772-z.

[16] M. Raj, S. Singh, K. Solanki, and R. Selvanambi, "An application to detect cyberbullying using machine learning and deep learning techniques," *SN Computer Science*, vol. 3, no. 5, Jul. 2022, doi: 10.1007/s42979-022-01308-5.

[17] S. M. Fati, A. Muneer, A. Alwadain, and A. O. Balogun, "Cyberbullying detection on Twitter using Deep Learning-Based attention mechanisms and continuous bag of words feature extraction," *Mathematics*, vol. 11, no. 16, p. 3567, Aug. 2023, doi: 10.3390/math11163567.

[18] V. L. Paruchuri and P. Rajesh, "CyberNet: a hybrid deep CNN with N-gram feature selection for cyberbullying detection in online social networks," *Evolutionary Intelligence (Print)*, vol. 16, no. 6, pp. 1935–1949, Sep. 2022, doi: 10.1007/s12065-022-00774-3.

[19] C. Van Hee et al., "Automatic detection of cyberbullying in social media text," *PloS One*, vol. 13, no. 10, p. e0203794, Oct. 2018, doi: 10.1371/journal.pone.0203794.

[20] Md. T. Hasan, Md. A. E. Hossain, Md. S. H. Mukta, A. Akter, M. Ahmed, and S. Islam, "A review on Deep-Learning-Based Cyberbullying Detection," *Future Internet*, vol. 15, no. 5, p. 179, May 2023, doi: 10.3390/fi15050179.

[21] P. Yi and A. Zubiaga, "Session-based cyberbullying detection in social media: A survey," *Online Social Networks and Media*, vol. 36, p. 100250, Jul. 2023, doi: 10.1016/j.osnem.2023.100250.

[22] Md. T. Hasan, Md. A. E. Hossain, Md. S. H. Mukta, A. Akter, M. Ahmed, and S. Islam, "A review on Deep-Learning-Based Cyberbullying Detection," *Future Internet*, vol. 15, no. 5, p. 179, May 2023, doi: 10.3390/fi15050179.

[23] N. Ananthi, "Cyber Bullying Types Datasets," IEEE Data Port, Aug. 31, 2021. https://ieee-dataport.org/documents/cyber-bullying-types-datasets

[24] S. K. Singh, K. Kumar, and B. Kumar, Sentiment Analysis of Twitter Data Using TF-IDF and Machine Learning Techniques. 2022. doi: 10.1109/com-it-con54601.2022.9850477.

[25] A. Singh, M. Jenamani, J. J. Thakkar, and Y. K. Dwivedi, "A text analytics framework for performance assessment and weakness detection from online reviews," *Journal of Global Information Management*, vol. 30, no. 8, pp. 1–26, Jul. 2022, doi: 10.4018/jgim.304069.

[26] A. Vaswani et al., "Attention is All you Need," arXiv (Cornell University), vol. 30, pp. 5998–6008, Jun. 2017.

[27] J. Vig, "Visualizing attention in Transformer-Based Language Representation models," arXiv (Cornell University), Apr. 2019.

[28] K. Zhang, P. Xu and J. Zhang, "Explainable AI in Deep Reinforcement Learning Models: A SHAP Method Applied in Power System Emergency Control," *2020 IEEE 4th Conference on Energy Internet and Energy System Integration (EI2)*, Wuhan, China, 2020, pp. 711-716, doi: 10.1109/EI250167.2020.9347147.