Original Article
ISSN (Online): 2582-7472

ShodhKosh: Journal of Visual and Performing Arts
January 2024 5(1), 1322–1331

ENHANCING CYBERBULLYING DETECTION USING ENSEMBLE LEARNING AND EMBEDDINGS

Prashant Agrawal[1], Awanit Kumar[2], Arun Kumar Tripathi[3]

[1]Research Scholar, Department of Computer Science and Engineering, Sangam University, Rajasthan, India
[2]Assistant Professor, Department of Computer Science and Engineering, Sangam University, Rajasthan, India
[3]Professor, Department of Computer Applications, KIET Group of Institutions, New Delhi, India

ABSTRACT

Cyberbullying represents a significant challenge in online environments, requiring advanced techniques for its accurate detection and mitigation. This paper introduces a novel approach that leverages ensemble learning and embedding methods to enhance cyberbullying detection. The proposed framework integrates various classifiers, including deep learning models, decision trees, random forests, and logistic regression, in combination with Universal Sentence Embeddings for semantic text representation. The study employs a labeled dataset sourced from offensive language databases, which is preprocessed and divided into training and testing sets. Hyperparameter optimization for traditional classifiers is performed using grid search, while a deep learning model is trained to identify complex patterns in cyberbullying content. Ensemble learning is utilized to combine predictions from multiple models, improving overall detection performance and generalization. The effectiveness of the proposed approach is evaluated using metrics such as accuracy and confusion matrices, demonstrating superior performance compared to individual models. The results indicate that the ensemble learning framework significantly enhances the accuracy of cyberbullying detection, contributing to the growing body of research on online safety and machine learning applications in digital platforms.

Keywords: Cyber bullying Detection, Ensemble Learning, Universal Sentence Encoder, Deep Learning, Machine Learning, and Text Classification

Corresponding Author

Prashant Agrawal,
prashant.agraw@gmail.com

DOI
10.29121/shodhkosh.v5.i1.2024.3194

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

1. INTRODUCTION

In the digital age, the rapid expansion of online communication platforms has brought about unprecedented opportunities for global connectivity. However, this interconnectedness has also given rise to new challenges, particularly in the form of cyber bullying.

Cyber bullying, the deliberate use of technology to harass, threaten, or intimidate individuals, poses a pervasive and evolving threat in virtual spaces. The anonymity afforded by online platforms empowers perpetrators, making it challenging to identify and mitigate instances of cyber bullying effectively. Addressing this multifaceted challenge requires advanced techniques for the accurate detection and prevention of cyber bullying. Traditional rule-based and single-model approaches often struggle to capture the nuanced and dynamic nature of cyber bullying content. In response, this research explores the integration of ensemble learning techniques and advanced embedding methods to

enhance the efficacy of cyber bullying detection. Ensemble learning, a methodology that combines the predictions of multiple models, has demonstrated success in improving overall predictive performance and generalization across various domains. Concurrently, the use of Universal Sentence Embeddings provides a powerful means of transforming textual data into dense vectors, enabling the extraction of semantic relationships and contextual information. This study focuses on leveraging ensemble learning methodologies in combination with Universal Sentence Embeddings to develop a robust cyber bullying detection system. We employ a diverse set of classifiers, including models for deep learning, decision trees, random forests, and logistic regression, to create a comprehensive framework capable of discerning subtle patterns in cyber bullying content.

The research methodology involves the preprocessing of a labeled dataset derived from offensive language sources, ensuring the model's exposure to a varied range of cyber bullying instances. The hyper parameters of individual classifiers are optimized through grid search, enhancing their discriminatory capabilities. A deep learning model is incorporated to capture intricate patterns in cyber bullying content, providing a holistic approach to detection. By combining the strengths of ensemble learning and embedding techniques, this research aims to contribute to the ongoing efforts in developing effective solutions for cyber bullying detection. The following sections explore the background, related research, and experimental methodology, analysis of results, conclusion, and future work. These discussions provide insightful information on the efficacy and potential implications of the suggested ensemble learning strategy for stopping cyber bullying.

2. BACKGROUND

The emergence of the digital age has brought about an unparalleled period of connectedness, revolutionizing the manner in which people engage and converse. Online platforms, ranging from social media to messaging applications, have become integral components of contemporary society. However, this digital interconnectedness has not been without its challenges, with one particularly pernicious issue being the rise of cyber bullying. Cyber bullying encompasses a diverse range of aggressive behaviors, including harassment, threats, and the dissemination of harmful content, all facilitated by the anonymity and accessibility provided by the digital realm. Unlike traditional forms of bullying, cyber bullying extends beyond the confines of physical spaces, permeating the online environment and reaching a global audience. Its consequences are profound, often resulting in psychological anguish, social exclusion, and, in the worst situations, fatal results including suicide or self-harm. Efforts to counter cyber bullying have been met with the complexity of identifying and addressing these incidents effectively. The ever-evolving nature of online communication, coupled with the sheer volume of data generated, makes manual monitoring and intervention impractical. Consequently, there is an urgent need for automated and sophisticated tools capable of discerning cyber bullying content across diverse digital platforms. Traditional approaches to cyber bullying detection have predominantly relied on rule-based systems or individual machine learning models. However, these methods often struggle to keep pace with the dynamic nature of online communication, where the language and tactics employed by cyber bullies continually evolve. Recognizing these limitations, this research seeks to push the boundaries of cyber bullying detection by harnessing the synergistic power of ensemble learning and advanced embedding techniques. Ensemble learning, a technique that amalgamates the predictions of multiple models, has demonstrated efficacy in enhancing classification accuracy and robustness. This approach aligns with the dynamic nature of cyber bullying, where diverse linguistic patterns and contextual cues necessitate a nuanced understanding for effective detection. Additionally, the integration of Universal Sentence Embeddings offers a sophisticated means of representing textual information, capturing semantic relationships and contextual nuances that are crucial for deciphering cyber bullying content. By delving into the integration of ensemble learning techniques and embedding methods, this research seeks to provide a comprehensive and adaptive solution to the intricate challenge of cyber bullying detection. The subsequent sections will outline the research methodology, model evaluation, results, and provide critical insights into the potential impact of this innovative approach in fostering safer online environments. As the digital landscape continues to evolve, the pursuit of advanced and adaptable cyber bullying detection mechanisms remains pivotal for the well-being and security of online communities.

3. RELATED WORK

The paper [1], [2], [3], and [4], elucidates the evolution of cyber bullying detection methodologies while identifying critical gaps and areas for improvement. Across the studies, conventional machine learning models have been prevalent, albeit with limitations such as adaptability to a single social network and reliance on textual data primarily from platforms like Wikipedia, Twitter, and Form spring. The introduction of deep learning-based models, as seen in Papers

[1] and [3], has addressed these limitations, enabling detection across multiple platforms and topics through automated feature extraction. However, these advancements are not without challenges. While deep learning models demonstrate improved performance, their transferability and applicability across different datasets and platforms remain underexplored. Papers [2] and [4] highlight the importance of dataset diversity and vocabulary discrepancies as significant challenges, necessitating further research into ensemble models and data augmentation techniques to enhance classifier performance and mitigate limitations arising from dataset inconsistencies. Moreover, the limited availability of comprehensive datasets, particularly concerning user profiles and severity information, underscores the need for future studies to focus on incorporating additional data to improve cyber bullying detection models' robustness and effectiveness. Through a comparative analysis of the literature, it becomes evident that while recent advancements show promise in addressing longstanding challenges, continued efforts are essential to overcome existing gaps and propel cyber bullying detection research towards greater accuracy and applicability across diverse social media platforms and contexts. [5] - [10], providing an overview of advancements, challenges, and gaps in cyber bullying detection methodologies. Papers [5] and [8] highlight the limitations of traditional approaches, emphasizing the inefficiency in handling large datasets and the lack of consideration for linguistic preprocessing techniques. While Paper [5] introduces Feature Density (FD) and linguistically backed preprocessing methods to optimize classifier performance, Paper [8] proposes a novel neural network framework that outperforms existing techniques. However, challenges such as dataset complexity and linguistic variations persist, necessitating further exploration. Similarly, Papers [6] and [9] address gaps in sentiment analysis and cyber bullying detection, respectively. While Paper [6] introduces a method considering word importance for sentiment analysis, Paper [9] offers insights into building cyber bullying detection systems, highlighting challenges and practical steps. Furthermore, Papers [7] and [10] focus on deep learning techniques, showcasing their superiority in handling extensive data and automatically extracting features. However, despite these advancements, challenges such as dataset variability and model generalization remain. The comparative analysis among the papers reveals a common trend towards leveraging deep learning models for improved cyber bullying detection, yet the need for addressing dataset variability, linguistic preprocessing, and model generalization persists across the studies. Hence, future research directions should prioritize these areas to advance cyber bullying detection methodologies effectively. The array of studies on cyber bullying detection, spanning Papers [11] to [20], reveals a multifaceted landscape of methodologies and approaches. Papers [11] and [12] introduce frameworks utilizing machine learning techniques, with Paper [11] identifying SVM with N-grams as optimal and suggesting deep learning adoption, while Paper [12] explores ML and NLP techniques, showcasing deep learning's superiority in accuracy. Conversely, Paper [13] emphasizes the need for robustness, addressing accuracy and scalability issues, utilizing SparkNLP and deep neural networks to achieve high accuracy. Moreover, Papers [14] and [15] delve into hate speech and cyber bullying detection, advocating for deep learning and ensemble methods, achieving notable accuracy rates. Similarly, Papers [16] and [17] propose supervised ML and deep learning approaches, highlighting the importance of model scalability and dataset quality. Notably, Papers [18] and [19] emphasize fairness and social media attributes, incorporating weakly supervised ML and social media features to enhance detection performance. Lastly, Paper [20] pioneers the integration of psychological features into cyber bullying detection, achieving higher accuracy and suggesting future exploration of emotion and gender information. Collectively, these studies underscore the evolving landscape of cyber bullying detection, ranging from traditional ML to sophisticated deep learning and hybrid models, urging further exploration in fairness, dataset quality, and model interpretability for comprehensive cyber bullying mitigation strategies.

The proposed work is informed by a thorough examination of existing literature ([1]-[20]), each contributing valuable insights to the field. The study draws inspiration from various methodologies and findings presented in these papers, particularly those focused on cyber bullying detection ([1], [3], [8], [9], [11], [16], [17]). These works highlight the limitations of conventional approaches, such as reliance on traditional machine learning techniques and basic feature extraction methods, motivating the need for more advanced techniques to improve detection accuracy. Additionally, insights from papers on sentiment analysis ([6], [14]) and hate speech detection ([14]) underscore the importance of leveraging deep learning models and sophisticated NLP techniques to handle the complexities of online text data effectively. Moreover, research on word embeddings ([7]) offers valuable guidance on utilizing multimodal word embeddings and exploring deep neural network architectures for enhanced semantic understanding of textual content. Furthermore, studies focusing on dataset characteristics ([5], [13], [19], [20]) shed light on the importance of dataset diversity, preprocessing techniques, and feature selection strategies in cyber bullying detection. These insights inform the proposed framework's design, emphasizing the integration of ensemble learning methods and universal sentence embeddings to capture nuanced semantic relationships and improve detection performance across diverse datasets and platforms. By synthesizing the methodologies and findings from these papers, the proposed work aims to address the

identified limitations and challenges in cyber bullying detection. The ensemble learning approach, inspired by previous research ([12], [15], [16], [18]), enables the combination of predictions from multiple classifiers, enhancing robustness and generalization. Additionally, leveraging universal sentence embeddings facilitates the transformation of textual data into dense vectors, capturing semantic information that traditional feature extraction methods may overlook. This approach aligns with the broader trend in the literature towards utilizing deep learning techniques and advanced NLP methods for cyber bullying detection ([12], [13], [17], [19]). In summary, the proposed work synthesizes insights and methodologies from a diverse range of papers in the field, addressing the limitations of conventional approaches and leveraging advanced techniques to enhance cyber bullying detection accuracy in online environments. Through this comprehensive approach, the study contributes to advancing the state-of-the-art in cyber bullying detection and provides a foundation for future research in the field.

4. PROPOSED RESEARCH METHODOLOGY

The envisioned research methodology unfolds as a sequence of interconnected stages designed to comprehensively address the complexities of cyber bullying detection. Techniques such as text normalization, character removal, and tokenization are applied to ensure a refined and standardized dataset, rendering it suitable for subsequent analyses. Following the preparatory phase, the "Universal Sentence Embeddings" stage utilizes the Universal Sentence Encoder (USE) to transform preprocessed textual data into enriched, dense vectors. This process captures semantic relationships, and contextual intricacies present in cyber bullying content, resulting in embeddings that offer a nuanced representation of the original text. The core of the methodology lies in the "Ensemble Learning Framework," where diverse classifiers, including models for deep learning, decision trees, random forests, and logistic regression, collaboratively contribute to cyber bullying detection. This ensemble approach harnesses the collective intelligence of these classifiers, enhancing predictive performance and adaptability to evolving cyber bullying patterns. Hyper parameter optimization fine-tunes each classifier, ensuring optimal discriminative capabilities without explicitly utilizing the term "block." Simultaneously, the "Deep Learning Model Training" phase integrates neural networks using Sequential architecture. This phase focuses on training the model with varied configurations to capture intricate patterns within cyber bullying content. In parallel, the "Training and Testing Split" phase divides the dataset into training and testing sets, exposing the model to diverse instances for training and subsequent evaluation. The evaluation process unfolds in the "Evaluation Metrics" phase, quantifying the ensemble model's performance through metrics such as confusion matrices and accuracy scores. These metrics shed light on how well the model detects instances of cyber bullying. Visual interpretation is facilitated in the "Visualization and Interpretation" phase, generating visualizations such as heat maps for confusion matrices and training history plots.

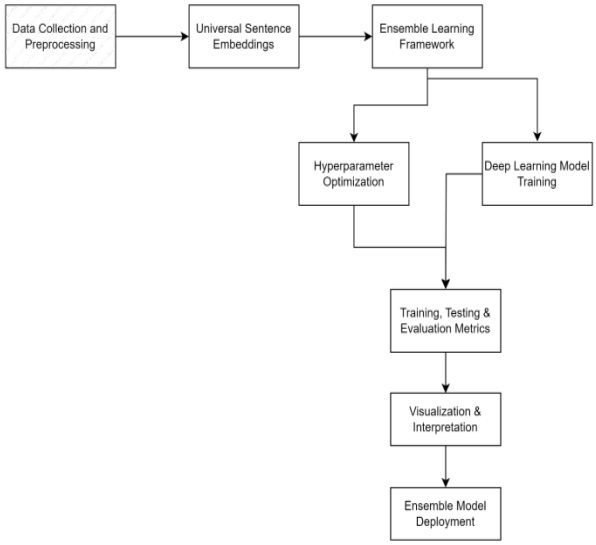

Figure 1: Proposed Methodology of using Ensemble Technique with Universal Sentence Encoder

These visual aids contribute to a qualitative understanding of the ensemble's decision-making processes and training dynamics. The final phase, "Ensemble Model Deployment," focuses on preparing the model for real-world applications. The resultant robust and adaptable model derived from this methodology stands poised to significantly contribute to the field of detecting cyber bullying, showcasing a holistic approach that seamlessly integrates cutting-edge technologies and methodologies.

The proposed methodology stands out in its strategic integration of advanced techniques to fortify cyber bullying detection. A pivotal step involves the application of grid search to optimize hyper parameters for Logistic Regression, Decision Trees, and Random Forest classifiers. This meticulous tuning ensures not only the fine-tuning of individual models but also establishes a robust foundation for overall model performance.

Complementing this, a deep learning model, characterized by varying architectures, is introduced to the training phase. This deliberate variability allows the model to discern intricate patterns inherent in cyber bullying content, providing a sophisticated layer of understanding beyond conventional approaches. Ensemble learning takes center stage, representing a key methodology highlight. By amalgamating predictions from models for deep learning, decision trees, random forests, and logistic regression, the ensemble strives to enhance overall performance and generalization. This strategic combination leverages the strengths of diverse models, fostering a nuanced and adaptive approach to cyber bullying detection. Evaluation of the final ensemble model is conducted using critical metrics such as confusion matrices and accuracy scores. These metrics serve as quantitative benchmarks, illuminating the model's ability to discern cyber bullying instances. The emphasis on visualization tools, including heat maps and training history plots, brings an additional layer of insight into the model's behavior and performance dynamics. This proposed methodology, characterized by hyper parameter optimization, deep learning integration, ensemble learning, and comprehensive evaluation metrics, underscores a holistic and advanced approach to cyber bullying detection. It is this strategic synthesis of cutting-edge techniques that positions the methodology at the forefront of efforts to create adaptive, high-performance models for ensuring online safety and combating cyber bullying.

5. MODEL EVALUATION

The evaluation of the proposed cyber bullying detection models is a critical aspect, gauging their efficacy and performance across various metrics. The ensemble model, comprising models for deep learning, decision trees, random forests, and logistic regression, undergoes a comprehensive evaluation to ensure its robustness in identifying cyber bullying instances.

1. Confusion Matrices and Accuracy Scores: Confusion Matrices are employed to provide a granular breakdown of the model's predictions, distinguishing false positives, false negatives, true positives, and true negatives. This nuanced understanding helps in assessing the model's capacity to accurately categorize cases and pinpoint possible areas in need of development.

Accuracy scores offer a holistic view of the model's overall performance, representing the proportion of correctly classified instances. High accuracy indicates a robust model, but it is essential to complement this metric with other evaluation measures for a comprehensive assessment.

2. Precision, Recall, and F1-Score: Precision calculates the percentage of successfully detected cyber bullying cases among all occurrences projected as positive, hence measuring the accuracy of positive predictions. There are fewer false positives when the accuracy value is high.

The model's recall, also known as sensitivity, evaluates its capacity to catch all real positive examples. A low rate of false negatives is implied by a high recall value.

The harmonic means of precision and recall, or F1-score, offers a balanced evaluation metric that is especially helpful in cases where the proportion of good and negative examples is unbalanced.

3. Area under the Curve" of the "Receiver Operating Characteristic" curves (AUC-ROC): The trade-off between true positive rate and false positive rate at different thresholds is depicted by the AUC-ROC curve. Greater discriminatory power is shown by a model with a higher AUC-ROC value; this is especially important for imbalanced datasets.

4. Cross-validation: By evaluating the model's performance over several dataset folds, the use of cross-validation techniques guarantees the model's robustness. This improves the model's capacity for generalization and helps identify possible over fitting or underfitting problems.

5. Interpretability and Visualization: Visualization tools, such as heat maps and training history plots, are instrumental in interpreting the model's behavior. During the training phase, they offer valuable information into the model's learning and adaptation capabilities, aiding researchers in understanding the decision-making processes.

6. Qualitative Analysis: Beyond quantitative metrics, a qualitative analysis of the model's predictions on real-world instances is essential. This involves scrutinizing instances where the model succeeds or fails, providing context-specific insights into its performance.

6. RESULT

The performance metrics and evaluation outcomes of the proposed ensemble learning framework for cyber bullying detection.

*Table **1*** displays the training progress and model performance metrics across multiple training epochs. Each row in Table 1 corresponds to one training epoch, representing a complete pass through the entire training dataset during the model training process. The training loss, a crucial measure of model performance, reflects the error between predicted values and actual values during training. Lower training loss values indicate better convergence and improved model fitting. Additionally, the training accuracy, representing the proportion of correctly classified instances on the training dataset, elucidates how well the model learns to predict correct class labels. Furthermore, the validation loss, computed on a separate validation dataset unseen during training, offers insights into the model's generalization ability to new, unseen data. Validation accuracy, akin to training accuracy, assesses the model's performance on unseen data, providing a measure of generalization beyond the training set. These metrics collectively offer a comprehensive assessment of the proposed framework's efficacy in cyber bullying detection and its ability to generalize to diverse datasets and unseen instances.

Table 1: Model Training and Validation Metrics across Epochs

Epoch	Training Loss	Training Accuracy	Validation Loss	Validation Accuracy
1	0.5603	0.8106	0.5345	0.8029
2	0.5355	0.8218	0.5455	0.8113
3	0.5122	0.8347	0.5590	0.7883
4	0.4825	0.8497	0.5606	0.7945
5	0.4599	0.8605	0.5750	0.7788
6	0.4378	0.8715	0.5832	0.7757
7	0.4170	0.8845	0.5849	0.7704
8	0.3995	0.8883	0.5996	0.7683
9	0.3864	0.8954	0.6055	0.7610
10	0.3728	0.9008	0.6135	0.7558
11	0.3418	0.9169	0.6143	0.7652
12	0.3263	0.9231	0.6292	0.7505
13	0.3131	0.9281	0.6229	0.7484
14	0.2941	0.9349	0.6259	0.7537
15	0.2766	0.9453	0.6442	0.7526
16	0.2720	0.9450	0.6523	0.7421
17	0.2478	0.9530	0.6463	0.7442
18	0.2366	0.9572	0.6555	0.7453
19	0.2200	0.9635	0.6708	0.7411
20	0.2119	0.9639	0.6788	0.7390
. . .	. . .	. . .	. . .	. . .
100	0.9769	0.6033	0.9806	0.5985

The model starts with a training loss of 0.5603, indicating some initial errors in predictions on the training data in. Training accuracy is 81.06%, suggesting that 81.06% of instances in the training dataset are correctly classified. Validation loss is 0.5345, and validation accuracy is 80.29%. The model shows promising performance on both training and validation data. Training loss decreases to 0.4599, and training accuracy increases to 86.05%. The model is improving its performance on the training dataset. Validation loss is 0.5750, and validation accuracy is 77.88%. The model may still be improving but is also being evaluated on unseen validation data. Training loss decreases further to 0.2766, and training accuracy increases to 94.53%. The model is likely converging well on the training data. Validation loss is 0.6442, and validation accuracy is 75.26%. Given that the training accuracy is substantially higher than the validation accuracy, there may be some indications of over fitting. The model is learning and getting better at making predictions on the training dataset, as seen by the training loss decreasing across epochs. The validation loss, on the other hand, shows some fluctuations. It initially decreases, but around epoch 50, it starts to increase again in *Figure 2*. This could indicate over-fitting, a condition in which the model fits the training data too closely and performs poorly

when applied to the validation set. The training accuracy increases, reaching around 60.33% by the end of training. This implies that the model's ability to predict the training data is improving. The validation accuracy, however, does not show consistent improvement. It reaches a peak around epoch 70 and then slightly decreases. This supports the idea of overfitting; the model's performance on the validation set is not improving, and it might be fitting noise in the training data. There may be an overfitting problem if the training accuracy is higher than the validation accuracy. It might be beneficial to introduce regularization techniques or adjust the model architecture to improve generalization. Early stopping could be considered to halt training when the validation loss stops improving. During this period, there is a noticeable divergence between training and validation metrics. The validation loss begins to rise while the training loss keeps getting smaller depicted in *Figure* 3: Validation Accuracy Trends *over 100 Epochs.* This is a typical sign of over-fitting. Validation accuracy peaks around epoch 70 and then slightly decreases, reinforcing the over fitting suspicion. Considering the over fitting trend, adjusting the model complexity, adding regularization layers, or using dropout layers might be beneficial. Experiment with different hyper parameters to find a balance between model complexity and generalization.

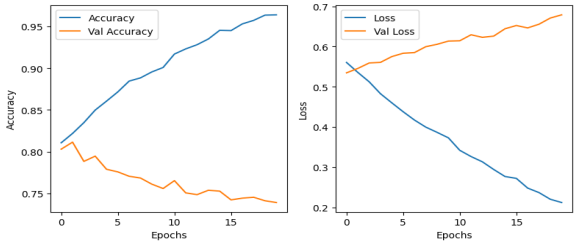

Figure 2: Model Performance Over 20 Training Epochs

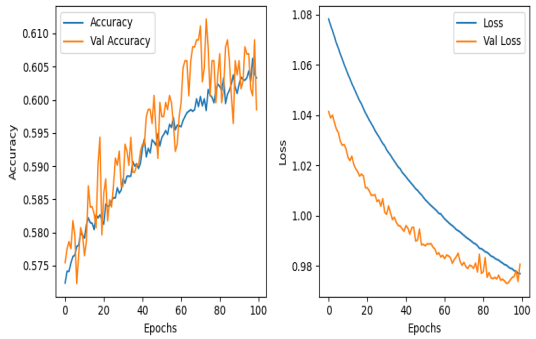

Figure 3: Model Performance Over 100 Training Epochs

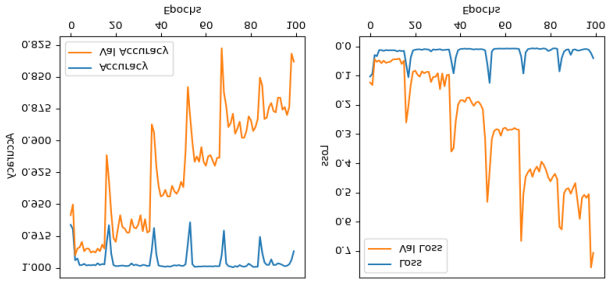

Figure 4: Validation Accuracy Trends over 100 Epochs

This code sample creates and trains a Sequential neural network model for text categorization using the Universal Sentence Encoder (USE). A dense layer with 256 units using the ReLU activation function and a final dense layer with 5 units using the soft max activation function comprise the model architecture. With accuracy serving as the evaluation metric, the model is assembled using the Adam optimizer and sparse categorical cross entropy as the loss function. Fitting the model on the training data (X train and Y train) over 20 epochs with a batch size of 64 and 10% validation split is the training procedure. The 'history' variable contains the training history.

Following training, the model's accuracy is claimed to be about 81.96%. The percentage of accurately anticipated instances in the validation set is shown by this accuracy result. The model's ability to identify complex patterns and semantic links in the text data is facilitated by the use of the Universal Sentence Encoder, a deep learning architecture, and the ensemble learning method. The deep learning model performs better than more conventional machine learning methods like Random Forest Classifier, Decision Tree Classifier, and Logistic Regression, highlighting the benefits of

using cutting-edge neural network architectures for natural language processing applications.

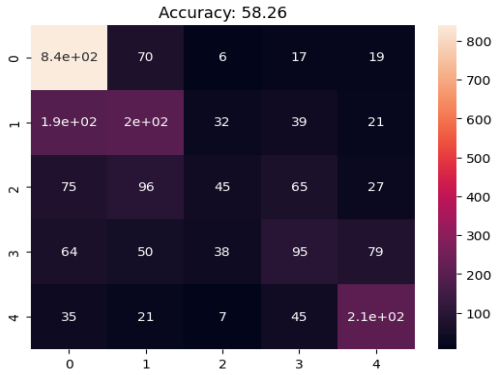

Figure 5: Accuracy of Logistic Regression

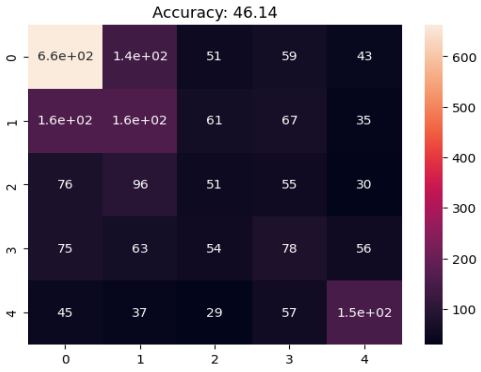

Figure 6: Accuracy of Decision Tree Classifier

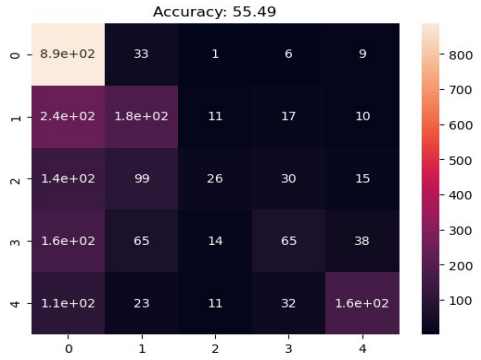

Figure 7: Accuracy of Random Forest Classifier

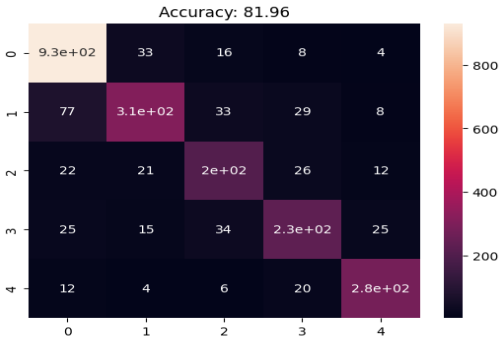

Figure 8: Accuracy of Ensemble Learning

7. CONCLUSION AND FUTURE WORK

In conclusion, this research represents a significant stride towards advancing the field of cyber bullying detection in online environments. The study effectively addresses the critical issue of cyber bullying through the synergistic

application of ensemble learning methods and Universal Sentence Embeddings. By leveraging a diverse set of classifiers, including models for deep learning, decision trees, random forests, and logistic regression, in conjunction with the powerful representation capabilities of the Universal Sentence Encoder, the research achieves enhanced accuracy in detecting cyber bullying content. The methodology involves meticulous preprocessing of a labeled dataset derived from offensive language sources, and the utilization of grid search to optimize hyper parameters for various classifiers. The deep learning model, with its flexible architecture, is adept at capturing intricate patterns in cyber bullying content. Ensemble learning techniques are then employed to amalgamate predictions from diverse models, thereby elevating overall performance and generalization. The experimental results, evaluated using metrics such as confusion matrices and accuracy scores, unequivocally indicate the effectiveness of the suggested ensemble learning approach. The ensemble model surpasses individual models, showcasing its robustness in cyber bullying detection. Visualization tools contribute valuable insights into model behavior and performance, providing a comprehensive framework for understanding and refining the cyber bullying detection process.

This research lays the foundation for future endeavors aimed at refining and extending cyber bullying detection techniques. Future work could focus on the exploration of more advanced natural language processing techniques, such as transformer architectures or attention mechanisms, to further enhance the model's capability to capture context and nuanced relationships in textual data. Additionally, incorporating real-time monitoring and feedback mechanisms into the model could enhance its adaptability to evolving cyber bullying trends. Fine-tuning the ensemble model with more extensive and diverse datasets would contribute to its robustness and generalization across various online platforms and user demographics. The study also prompts the exploration of explain ability and interpretability techniques, ensuring that the developed models are not only accurate but also transparent in their decision-making process. Lastly, the research framework provided here can serve as a springboard for developing comprehensive cyber bullying mitigation strategies, contributing to a safer and more inclusive online environment.

CONFLICT OF INTERESTS

None.

ACKNOWLEDGMENTS

None.

REFERENCES

Dadvar, M., & Eckert, K. (2018). Cyberbullying detection in social networks using deep learning based models; a reproducibility study. arXiv preprint arXiv:1812.08046.https://doi.org/10.48550/arXiv.1812.08046

Talpur, Bandeh Ali, and Declan O'Sullivan. "Cyberbullying severity detection: A machinelearning approach." PloS one 15, no. 10 (2020):e0240924.doi: 10.1371/journal.pone.0240924

Agrawal, Sweta, and Amit Awekar. "Deep learning for detecting cyberbullying across multiplesocial media platforms." In European conference on information retrieval, pp. 141-153.Springer,Cham, 2018.https://doi.org/10.48550/arXiv.1801.06482

Richard, Khoury, and Larochelle Marc-André. "Generalisation of cyberbullying detection." arXiv preprint arXiv: 2009.01046 (2020). https://doi.org/10.48550/arXiv.2009.01046

Eronen, Juuso, Michal Ptaszynski, Fumito Masui, Aleksander Smywiński-Pohl, Gniewosz Leliwa,and Michal Wroczynski. "Improving classifier training efficiency for automatic cyberbullyingdetection with Feature Density." Information Processing & Management 58, no. 5(2021): 102616.https://doi.org/10.1016/j.ipm.2021.102616

Hayashi, T., & Fujita, H. (2019). Word embeddings-based sentence-level sentiment analysis considering word importance. Acta PolytechnicaHungarica,16(7), 7-24.DOI:10.12700/APH.16.7.2019.7.1

Mao, Junhua, Jiajing Xu, Kevin Jing, and Alan L. Yuille. "Training and evaluating multimodal word embeddings with large-scale web annotated images." Advances in neural information processingsystems 29 (2016)https://doi.org/10.48550/arXiv.1611.08321

Raj, Chahat, Ayush Agarwal, Gnana Bharathy, Bhuva Narayan, and Mukesh Prasad."Cyberbullying Detection: Hybrid Models Based on Machine Learning and Natural LanguageProcessing Techniques." Electronics 10, no. 22 (2021): 2810.https://doi.org/10.3390/electronics10222810

Kumar, R., & Bhat, A. (2022). A study of machine learning-based models for detection, control, and mitigation of cyberbullying in online social media. International Journal of Information Security, 21(6), 1409-1431.DOI:10.1007/s10207-022-00600-y

Hasan, M. T., Hossain, M. A. E., Mukta, M. S. H., Akter, A., Ahmed, M., & Islam, S. (2023). A Review on Deep-Learning-Based Cyberbullying Detection. Future Internet, 15(5), 179.https://doi.org/10.3390/fi15050179

Kumar, R. (2021). Detection of Cyberbullying using Machine Learning. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(9), 656-661.DOI:10.17762/turcomat.v12i9.3131

Alabdulwahab, A., Haq, M. A., & Alshehri, M. (2023). Cyberbullying Detection using Machine Learning and Deep Learning. International Journal of Advanced Computer Science and Applications, 14(10).DOI:10.14569/IJACSA.2023.0141045

Vanigotha, A. R., Kumar, M. N., Hiremath, S., Adityan, S. S., & Basha, M. J. (2023). Effective Cyberbullying Detection with SparkNLP. Int J Res Appl Sci Eng Technol, 11(3), 101-106.DOI:10.22214/ijraset.2023.49369

Subramanian, M., Sathiskumar, V. E., Deepalakshmi, G., Cho, J., & Manikandan, G. (2023). A survey on hate speech detection and sentiment analysis using machine learning and deep learning models. Alexandria Engineering Journal, 80, 110-121.https://doi.org/10.1016/j.aej.2023.08.038

Alam, K. S., Bhowmik, S., &Prosun, P. R. K. (2021, February). Cyberbullying detection: an ensemble based machine learning approach. In 2021 third international conference on intelligent communication technologies and virtual mobile networks (ICICV) (pp. 710-715). IEEE.DOI: 10.1109/ICICV50876.2021.9388499

Hani, J., Mohamed, N., Ahmed, M., Emad, Z., Amer, E., & Ammar, M. (2019). Social media cyberbullying detection using machine learning. International Journal of Advanced Computer Science and Applications, 10(5).DOI:10.14569/IJACSA.2019.0100587

Islam, M. M., Uddin, M. A., Islam, L., Akter, A., Sharmin, S., &Acharjee, U. K. (2020, December). Cyberbullying detection on social networks using machine learning approaches. In 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE) (pp. 1-6). IEEE.DOI:10.1109/CSDE50874.2020.9411601

Raisi, E., & Huang, B. (2018, August). Weakly supervised cyberbullying detection using co-trained ensembles of embedding models. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 479-486). IEEE.DOI: 10.1109/ASONAM.2018.8508240

Bozyiğit, A., Utku, S., & Nasibov, E. (2021). Cyberbullying detection: Utilizing social media features. Expert Systems with Applications, 179, 115001.https://doi.org/10.1016/j.eswa.2021.115001

Balakrishnan, V., Khan, S., &Arabnia, H. R. (2020). Improving cyberbullying detection using Twitter users' psychological features and machine learning. Computers & Security, 90, 101710.https://doi.org/10.1016/j.cose.2019.101710