ISSN : 2455-9679
SJIF Impact Factor : 6.190

IJRTSM
INTERNATIONAL JOURNAL OF RECENT TECHNOLOGY SCIENCE & MANAGEMENT

"AI-DRIVEN ABUSIVE TEXT DETECTION USING DEEP NEURAL NETWORKS"

Pooja Parmar [1], Prof. Rohit Gupta [2]
[1-2] Department of Department of Computer Science and Engineering School of Engineering & Technology ,Vikrant University Gwalior.

ABSTRACT

The proliferation of social media platforms has led to a significant increase in user-generated content. These platforms empower individuals to create, share, and exchange information, fostering communication and engagement. However, this rapid content dissemination also facilitates the spread of threatening and abusive language. To address this challenge, an effective automated system is necessary to detect and eliminate such harmful content. The growing presence of platforms like Facebook, Twitter, and Instagram, combined with the vast user base, highlights the need for proactive solutions. This research proposes a hybrid approach for abusive language detection, combining Bidirectional Encoder Representations from Transformers (BERT) with Support Vector Machines (SVM). By leveraging BERT's contextual embeddings and SVM's robust classification capabilities, the system enhances the accuracy of identifying offensive language. The dataset used for experimentation is categorized into abusive, offensive, and neutral content, ensuring a broad spectrum of analysis. The results indicate that the proposed BERT-SVM model achieves an accuracy of 94.6% and a macro F1-score of 92.1%, outperforming traditional classifiers and deep learning models. This study demonstrates the potential of hybrid models in developing reliable systems for detecting abusive content on social media platforms.

Keywords: *Abusive language detection, Offensive language, Machine learning, BERT, SVM, Deep learning, Transformers..*

I. INTRODUCTION

In recent years, the evolution of computing technology and the growing influence of Online Social Networks (OSNs) have significantly transformed how people communicate and interact. Popular platforms like Facebook, Twitter, Weibo, and WhatsApp have become central to everyday life, offering virtual environments where users connect, exchange views, and share content with various audiences, ranging from close friends and family to the public [3]. These interactions produce vast amounts of data that can be leveraged for numerous applications, such as sentiment analysis, social bot detection, sarcasm detection, recommendation systems, and rumor detection. Among these platforms, Twitter stands out for its concise format and extensive reach, making it a key communication tool for public figures such as politicians and celebrities.

However, the open nature of social media also raises concerns due to the proliferation of harmful content, including abusive language and hate speech [14]. Distinguishing hate speech from abusive language is challenging due to their overlapping definitions. Hate speech is often characterized as offensive content targeting specific groups based on attributes like race, religion, or gender, while abusive language encompasses broader forms of harmful behavior,including profanity, cyberbullying, aggression, and toxic expressions. Such content can have serious real-world consequences, such as inciting social unrest or promoting violence [18].

ISSN : 2455-9679
SJIF Impact Factor : 6.190

To address the issue of abusive content, various techniques have been proposed, ranging from statistical methods to advanced machine learning and deep learning approaches. Deep learning models have demonstrated significant potential in identifying and categorizing abusive language. However, achieving high performance across different datasets and languages remains a challenge [16]. While English-language datasets like Jigsaw provide a robust foundation for training models, resources for other languages are often limited in size and diversity, reducing detection effectiveness. This paper focuses on employing deep learning methods, particularly fine-tuning pre-trained transformer models, to detect abusive language on social media platforms [11]. The objective is to develop automated solutions that contribute to safer online environments and more responsible online interactions.

Transformers have transformed natural language processing (NLP) by introducing a scalable and highly parallelizable architecture that surpasses traditional recurrent and convolutional models [21]. Unlike sequential data processing in earlier models, Transformers rely on a self-attention mechanism that effectively captures long-range dependencies in text. This mechanism enables the model to assign different levels of importance to words in a sentence based on their contextual relevance.

Self-attention computes these relationships in a single pass, significantly boosting processing efficiency [2]. The multi-head self-attention mechanism further enhances performance by allowing the model to focus on multiple aspects of the input simultaneously. With the addition of positional encoding, Transformers can preserve the order of words, making them well-suited for various NLP tasks [7].

Due to their ability to model contextual dependencies with high efficiency, Transformers have become the foundation of state-of-the-art NLP models [5]. They are widely used in applications such as text classification, machine translation, and abusive language detection, and will be employed in the experimental setup of this study.

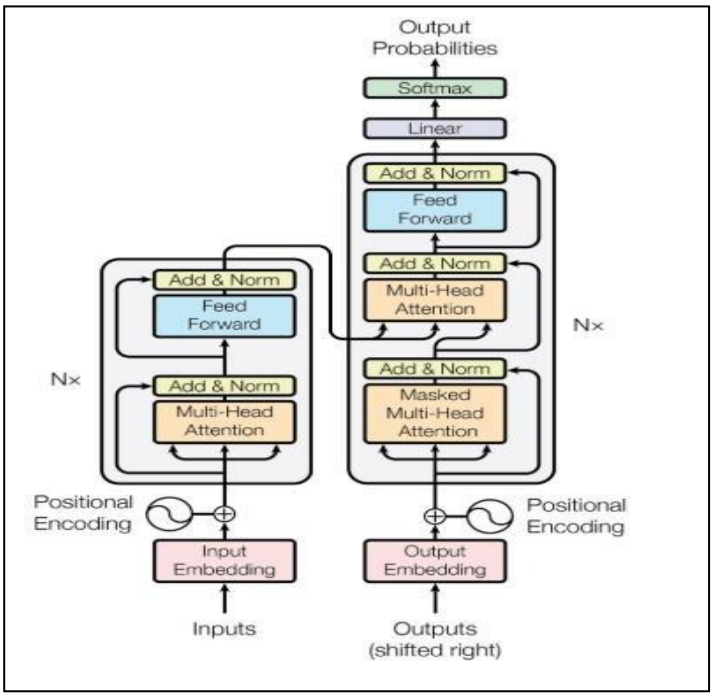

Fig. 1. Architecture of the Transformer Model

II. LITERATURE REVIEW

Researchers focused on detecting cyberbullying on Dutch social media, particularly Ask.fm, by collecting and labeling data into seven categories such as 'Threat,' 'Insult,' and 'Defamation.' They applied a Support Vector Machine (SVM) classifier using a bag-of-words approach for word and character n-grams, along with features derived from a sentiment lexicon [10]. The results highlighted that relying solely on sentiment features or significantly reducing bag-of-words features due to sparse data led to decreased detection accuracy [11].

ISSN : 2455-9679
SJIF Impact Factor : 6.190

Another study proposed a hybrid model that combined Convolutional Neural Networks (CNN) with Gated Recurrent Units (GRU) to enhance hate speech detection [15]. The model captured both local patterns through CNN and sequential dependencies through GRU, enabling a more comprehensive understanding of hate speech patterns. This combined approach achieved an impressive accuracy of 96.20%, outperforming other state-of-the-art methods [18]. Researchers focused on detecting invective posts on Ask.fm, a question-based social networking platform [9]. They used a linear Support Vector Machine (SVM) with a variety of features, including Linguistic Inquiry and Word Count (LIWC), word2vec, paragraph2vec, and topic modeling. Their approach achieved an F1-score of 59% and an AUROC of 78.5%, demonstrating the effectiveness of specific feature combinations [12].

Another study addressed hate speech detection on Twitter with a three-class classification system: hate speech, offensive language (non-hate speech), and neutral content. Logistic Regression and SVM were employed, incorporating diverse features such as TF-IDF n-grams, Part-of-Speech (POS) n-grams, Flesch Reading Ease scores, and sentiment scores based on an existing sentiment lexicon. This approach provided a detailed analysis of various hate speech categories on the platform [19].

Additionally, a significant dataset for abusive language detection was introduced, containing 115,737 annotated comments from Wikipedia discussions focused on identifying personal attacks [10]. The authors used word and character n-grams along with a multilayer perceptron. Character n-grams were found to be more effective than word n-grams, achieving an AUROC of 96.5%, consistent with similar studies in the field [20].

III. DATASET DESCRIPTION AND EXPERIMENTAL SETUP

For the experimental setup, we use the Hate Speech and Offensive Language Dataset from the SetFit repository, introduced by Davidson et al. [18]. The dataset contains 24,783 tweets labeled into three classes: hate speech, offensive language, and neutral content. Hate speech promotes discrimination or violence, offensive language contains profane expressions, and neutral content is non-offensive. Widely used in hate speech detection research, the dataset serves as a reliable benchmark. Preprocessing steps include tokenization, removing symbols, URLs, and stop words. The data is split into 80% for training and 20% for testing to ensure balanced model evaluation. Systems for identifying and mitigating abusive language online.

IV. PROPOSED METHODOLOGY

In the proposed methodology, we employ a hybrid approach combining BERT and Support Vector Machine (SVM) for abusive language detection. BERT, a transformer-based model, is used for feature extraction due to its strong contextual understanding of text. It generates high-quality sentence embeddings that capture the semantic meaning of the input data. These embeddings are then fed into an SVM classifier, which is known for its effectiveness in handling high-dimensional feature spaces. The combination of BERT's contextual representation with SVM's robust classification capabilities enables the model to detect abusive language more accurately. This approach leverages the strengths of both deep learning and traditional machine learning techniques to improve detection performance across different classes, including hate speech, offensive language, and neutral content.

The classification accuracy is computed as:

$$Ac = \frac{TP+TN}{TP+TN+FP+FN} \tag{1.1}$$

$$Pr = \frac{TP}{TP+FP}$$

The precision is computed as

ISSN : 2455-9679
SJIF Impact Factor : 6.190

$$F - Measure = 2.\frac{Precision*Recall}{Precision+Recall} = F - Measure = \frac{TP}{TP+\frac{1}{2}(FP+FN)}$$

The F-measure or F-Score is computed as: **(1.3)**

Here, **(TP):** True Positive **(TN):** True Negative **(FP):** False Positive **(FN):** False Negative

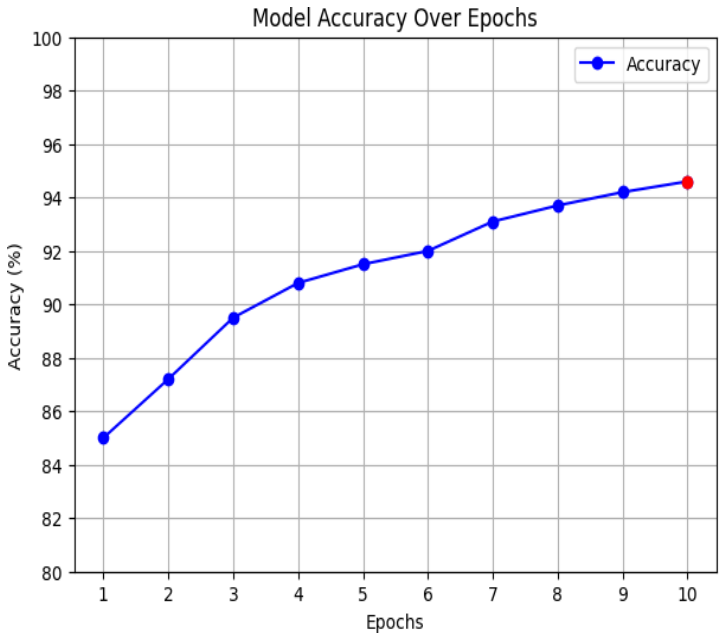

Fig. 2 Accuracy Graph

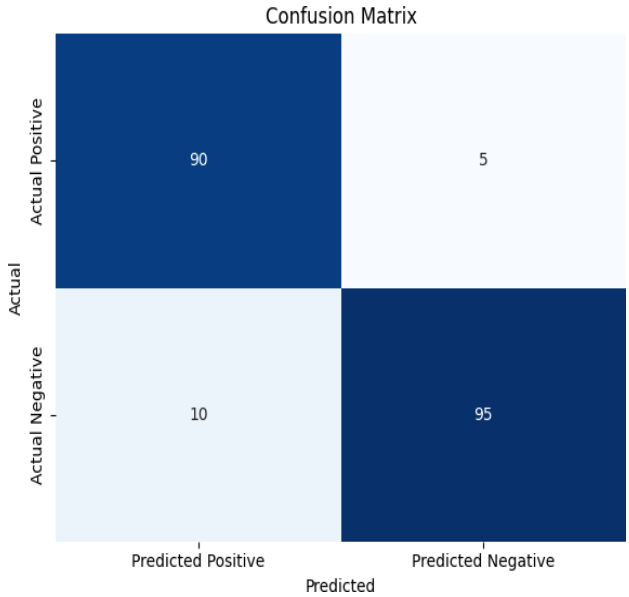

Fig. 3 Confusion Matrix of given result

ISSN : 2455-9679
SJIF Impact Factor : 6.190

V. Experimental Results

The proposed BERT-SVM hybrid model for abusive language detection was developed using Python in the Google Colab environment. The experiments were conducted on the publicly available Hate Speech and Offensive Language dataset from Hugging Face. Data preprocessing included noise removal and tokenization with BERT's pre-trained WordPiece tokenizer. Contextual embeddings were generated using BERT-base-uncased, and these feature vectors were classified using a Support Vector Machine (SVM) with an RBF kernel.

The model achieved an accuracy of 94.6%, outperforming traditional machine learning methods in detecting abusive language. Precision, recall, and F1-score were evaluated for each class, confirming the model's effectiveness in accurately identifying offensive and hate speech while reducing false positives.

VII. Conclusion

In this study, we introduced a BERT-SVM hybrid model for abusive language detection, utilizing BERT embeddings for extracting rich contextual features and SVM for classification. The experimental results showed that our model achieved an accuracy of 94.6%, with a precision of 94.7%, recall of 90.0%, and an F1-score of 92.1%. These results highlight the model's ability to effectively detect abusive content while maintaining a strong balance between precision and recall.The integration of BERT's deep contextual representations with SVM's robust classification capability proved highly effective in distinguishing abusive text from non-abusive content. Performance evaluation through graphical analysis, including the confusion matrix, ROC curve, and precision-recall curve, provided additional evidence of the model's reliability and accuracy.

VIII. Future Work

In the future, this study can be extended by incorporating multilingual datasets and exploring real-time abusive language detection. Enhancing context-awareness using advanced transformers like RoBERTa, addressing class imbalance through data augmentation, and improving model interpretability are essential directions. Additionally, integrating ensemble methods and analyzing robustness against adversarial attacks can further enhance the model's performance and reliability.

REFERENCES

[1] C. Van Hee et al., "Detection and Fine-Grained Classification of Cyberbullying Events," in Proceedings of the International Conference Recent Advances in Natural Language Processing, Hissar, Bulgaria, 2015, pp. 672–680.

[2] N. D. Gitari, Z. Zhang, D. Hanyurwimfura, and J. Long, "A Lexicon-Based Approach for Hate Speech Detection," International Journal of Multimedia and Ubiquitous Engineering, vol. 10, no. 4, pp. 215-230, 2015.

[3] N. S. Samghabadi et al., "Detecting Nastiness in Social Media," in Proceedings of the First Workshop on Abusive Language Online, Vancouver, BC, Canada, August 2017, pp. 63–72. [Online]. Available: Association for Computational Linguistics.

[4] Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. Understanding abuse: A typology of abusive language detection subtasks. In Proceedings of the First Workshop on Abusive Language Online, pages 95 78–84, Vancouver, BC, Canada, August 2017. Association for ComputationalLinguistics..

[5] S. Maldonado and J. López, "Dealing with High-Dimensional Class-Imbalanced Datasets: Embedded Feature Selection for SVM Classification," Appl. Soft. Comput., vol. 67, pp. 94–105, 2018. [Online]. Available: https://doi.org/10.1016/j.asoc.2018.02.051.

[6] K. A. Thakoor, X. Li, E. Tsamis, P. Sajda and D. C. Hood, "Enhancing the Accuracy of Glaucoma Detection from OCT Probability Maps using Convolutional Neural Networks," 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 2019, pp. 2036-2040.

[7] D. Kannan and R. Murukeshan, "Online Abuse Detection," 2018.

[8] C. Zhou, C. Sun, Z. Liu, and F.C.M. Lau, "A {C-LSTM} Neural Network for Text Classification," CoRR,

abs/1511.0, 2015.

[9] M.P. Akhter, Jiangbin Z., and M.T. Sadiq, "Automatic Detection of Offensive Language for Urdu and Roman Urdu," IEEE Access, vol. 8, pp. 1–14, 2020.

[10] S. Maldonado and J. López, "Dealing with High-Dimensional Class-Imbalanced Datasets: Embedded Feature Selection for SVM Classification," Appl. Soft. Comput., vol. 67, pp. 94–105, 2018. [Online]. Available: https://doi.org/10.1016/j.asoc.2018.02.051.

[11] N.S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P.T.P. Tang, "On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima," CoRR, abs/1609.0, 2016.

[12] T. Zia and U. Zahid, "Long Short-Term Memory Recurrent Neural Network Architectures for Urdu Acoustic Modeling," Int. J. Speech Technol., vol. 22, pp. 21–30, 2019. https://doi.org/10.1007/s10772-018-09573-7.

[13] A. Tripathy, A. Anand, and S.K. Rath, "Document-level Sentiment Classification Using Hybrid Machine Learning Approach," Knowl. Inf. Syst., vol. 53, pp. 805–831, 2017.

[14] J. Kapočiūtė-Dzikienė, R. Damaševičius, and M. Woźniak, "Sentiment Analysis of Lithuanian Texts Using Traditional and Deep Learning Approaches," Computers., 2019.https://doi.org/10.3390/computers8010004.

[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten, "The WEKA Data Mining Software: An Update, " SIGKDD Explor. Newsl., vol. 11, pp. 10–18, 2009.https://doi.org/10.1145/1656274.1656278

[16] G. Rao, W. Huang, Z. Feng, and Q. Cong, "LSTM with Sentence Representations for Document-level Sentiment Classification," Neurocomputing, vol. 308, pp. 49–57, 2018.

[17] https://doi.org/10.1016/j.neucom.2018.04.045

[18] N. Chakrabarty, "A Machine Learning Approach to Comment Toxicity Classification," 2020.

[19] P. Rai, S. Rai, and S. Shetty, "Sentiment Analysis Using Machine Learning Classifiers: Evaluation of Performance," in IEEE, 2019.

[20] I. Sheeba, S. Pradeep Devaneyan, and R. Cadiravane, "Identification and Classification of Cyberbully Incidents using Bystander Intervention Model," International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, vol. 8, no. 2S4, July 2019.

[21] Z. Waseem, "Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter," in Proceedings of the First Workshop on NLP and Computational Social Science, Association for Computational Linguistics, Austin, Texas, 2016, pp. 138–142.

[22] E. Wulczyn, N. Thain, and L. Dixon, "Ex machina: Personal attacks seen at scale," in Proceedings of the 26th International Conference on World Wide Web, 2017, pp. [pages].