International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 02 | Feb - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

DetectNet: A Hybrid and Rule-Based and BERT – Powered Cyberbullying Detection System

Dewansh Dewangan[1], Namrata Bahadur[2], Chaitanya Jadhav[3], Rohit Nevare[4]

Department of Information Technology, Sinhgad Academy of Engineering, Pune

ABSTRACT:
In online communication platforms, cyberbullying is a serious issue due to the limits of traditional keyword detection methods and the lack of context in automated systems. In this paper, we present Detect-Net, a cyberbullying detection framework that combines rule-based language analysis with BERT-based deep learning. The system detects both clear and subtle bullying behaviors and sorts them into various categories. Detect-Net also offers clear outputs to improve trust and transparency, making it suitable for real-time moderation and online safety.

Key Words: Cyberbullying Detection, Natural Language Processing, BERT, Hybrid Classification, Explainable AI, Text Classification, Social Media Safety.

1.INTRODUCTION

The issue of cyberbullying is continuing to increase with the rapid proliferation of social media and online communication channels. Many online platforms have implemented filters based on keywords to detect abusive messages; however, traditional keyword filtering methods are often not able to accurately identify context-based forms of abuse, such as sarcasm, or forms of implicit bullying. As technology continues to advance, systems that detect inappropriate behaviour through text will need to be more intelligent and contextually aware.

There have been many methods created over the past couple of years for identifying cyber-bullying via the use of machine learning language processing (NLP). Although many of the established detection methods have improved their ability to detect, many continue the process of cyberbyllying.

II. PRIOR RESEARCH

The quick spread of harmful content on social media platforms has drawn a lot of research interest in cyberbullying detection. The majority of early methods were rule-based and lexicon-based, identifying harmful messages using manually selected dictionaries of offensive terms and syntactic patterns. High false-negative rates resulted from these systems' poor generalisation, which failed to capture context, sarcasm, and changing slang. However, they demonstrated high precision for obviously offensive content.

Then, using surface-level features like n-grams, term frequency–inverse document frequency (TF-IDF), and sentiment scores, conventional machine learning (ML) classifiers like Support Vector Machines (SVM), Naïve Bayes, and Random Forest were investigated. Although these models were more flexible and scalable than rule-based methods, they still needed a lot of feature engineering and lacked deep semantic understanding.

Cyberbullying detection advanced significantly with the introduction of deep learning models. By learning hierarchical representations from text, recurrent neural networks (RNNs), convolutional neural networks (CNNs), and long short-term memory (LSTM) networks showed superior performance. By bidirectionally capturing contextual dependencies, transformer-based architectures—specifically, BERT (Bidirectional Encoder Representations from Transformers)—have recently produced state-of-the-art results in abusive language detection. However, in situations where surface cues are crucial, models that are solely based on data may incorrectly classify content.

DetectNet was developed in response to the proposal of hybrid models that combine contextual

representation learning and rule-based heuristics in order to overcome these limitations. For reliable cyberbullying detection, this integration aims to combine the semantic power of BERT with the accuracy of rules.

III. PROJECT OBJECTIVE

DetectNet is a robust cyberbullying detection system that uses a hybrid rule-based and BERT-powered deep learning architecture to effectively identify abusive, harmful, and offensive online content. Its main goal is to design and develop this system. The following are the specific goals of the suggested work:

1) **To Create a Hybrid Detection Framework:** To get around the drawbacks of stand-alone methods, the project intends to combine rule-based linguistic heuristics with BERT-based contextual language modelling. While the BERT model captures implicit, contextual, and semantically complex bullying expressions, the rule-based module concentrates on identifying explicit cyberbullying patterns like hate speech, profanity, and direct harassment.

2)**To Increase Contextual Understanding and Detection Accuracy:** The ability to recognise subtle forms of cyberbullying, such as sarcasm, indirect insults, and contextual abuse, is another important goal in improving overall classification performance. The system aims to increase recall and decrease false negatives while maintaining precision by leveraging BERT's bidirectional attention mechanism.

3)**ToAssure Scalability and Adaptability:** The project aims for a modular and scalable architecture that makes it simple to update offensive rule sets, lexicons, and BERT model retraining.

4)**To Support Real-World Deployment and Online Safety:** Lastly, the goal is to create a workable system that can be implemented in real-world settings like social media moderation tools and classrooms, promoting safer online communication and efficient abuse prevention.

IV. OVERVIEW AND FEATURES

The proposed **DetectNet** system is a hybrid cyberbullying detection framework designed to accurately identify abusive and harmful online content by combining rule-based intelligence with BERT-powered contextual understanding. The key features of DetectNet are described as follows:

1)Hybrid Detection Architecture:
DetectNet integrates a deterministic rule-based module with a transformer-based BERT model, enabling the system to detect both explicit and implicit forms of cyberbullying with higher reliability than single-model approaches.

2)Rule-Based Linguistic Analysis:
The rule-based component employs curated lexicons, pattern matching, and syntactic rules to capture profanity, hate speech, threats, and repeated harassment, ensuring high precision in clearly abusive cases.

3)**Context-Aware BERT:** The BERT module also analyzes bidirectional context and semantic relationships. This is helpful in properly interpreting expressions of sarcasm, as well as indirect insults or context-dependent bullying.

4)**Sequential Decision Fusion Mechanism:** The strategy follows a decision fusion concept, where the rules are used in conjunction with the BERT model for final classification, so that false positives and false negatives are avoided.

5) **Noise and Informal Text Handling:** It is also robust against any noisy input, such as slang, spelling mistakes, abbreviations, emoticons, or grammar commonly found on social media sites.

6)**Multi-Class Cyberbullying:** The model supports classification in multiple categories, such as bullying, harassment, hate, and non-offensive content.

7) **Scalability and Modular Design:**
This modular architecture enables rule set, lexicon, and BERT model updates to be handled independently of one another, scaling the system to keep up with evolving online language.

8) **Capability of Dataset Generalization:**
It is designed for generalization across datasets and platforms, thus reducing dataset-specific bias while improving cross-domain performance.

9**. Real-Time Processing Capability:** DetectNet has efficient inference and is therefore suitable for near real-time content moderation tasks within large-scale online systems.

10**) Support for AI - Ethics and Responsibility:** The framework ensures fairness and responsible detection through minimum bias, interpretable rules that maintain transparency, and support for human-in-the-

loop moderation.

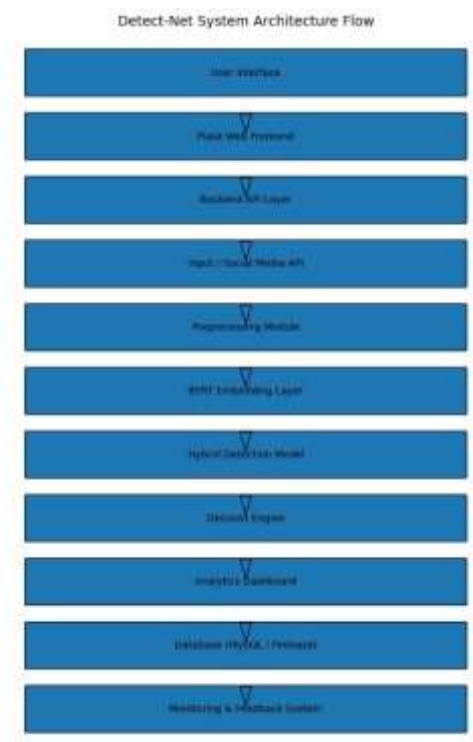

Figure 1 : Architecture Diagram Flow

The architecture of DetectNet is based on a layered and modular approach to ensure robustness, interpretability, and detection accuracy. First is the text preprocessing layer, which normalizes, tokenizes, removes noise, and addresses slang, emojis, and misspellings that are very common on social media platforms. The output is then fed into parallel paths comprising a rule-based analysis engine and a BERT module for a deep learning approach, respectively. The output from each module is then fused together using a decision fusion layer before a final output is generated. In the BERT module, a deep learning approach is used to generate a classification output based on contextual embeddings. The second step is a rule-based module based on a predefined dictionary, syntax patterns, and heuristics to detect explicit abusive language that indicates cyberbullying. These two modules run concurrently, and their results are then passed through a decision fusion module to produce a unified output.

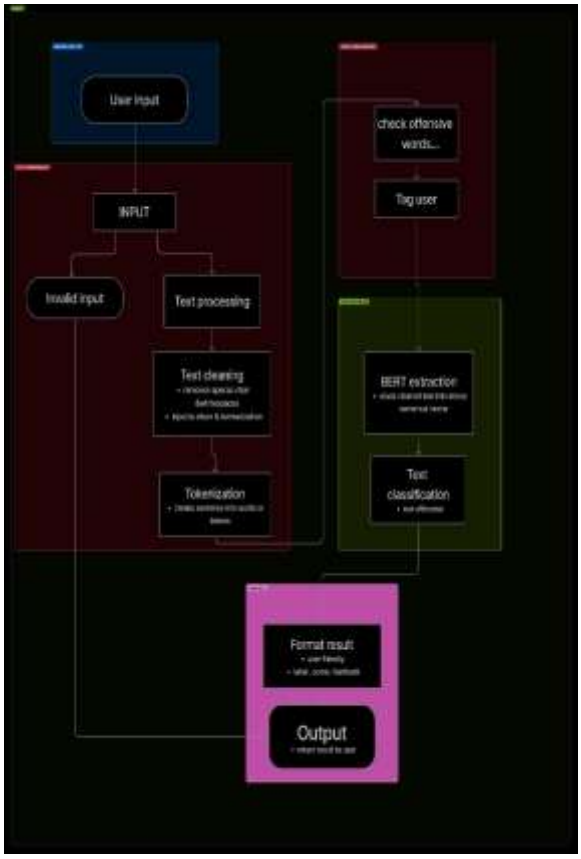

Figure 2: System Class Diagram Flow

The system class diagram for Detect Net specifies the modular and layered architecture that clearly conveys responsibilities and data flow among the different components. At the top level, the User Interface class is the point of interaction, receiving raw text input from users or other platforms. A User Interface class should pass this to the Input Handler, validating that data and then acting upon it down into pre-processing.

The Preprocessor class deals with the normalization of text, tokenization, removal of noise, emoji handling, and normalizing slang. The cleaned text was then dispatched parallel to the two core analytical classes: Rule Based Detector and BERT Classifier.

The Rule Based Detector applies predefined lexicons, pattern matching rules, and heuristic conditions for the explicit detection of cyberbullying indicators. At the same time, the BERT Classifier encodes the text with contextual embeddings and analyzes semantically to identify implicit or context-dependent bullying. Outputs of the two detection classes each are then fed into the Decision Fusion class, which fuses the rule-based confidence scores with the BERT prediction probabilities to come up with a final decision. The

Result Manager class formats the outcome of classification and stores it; results can be logged optionally for evaluation or auditing purposes. Finally, the processed output is returned to the User Interface for display moderation action.

V. SEQUENCE FLOW

The depicted sequence diagram describes the interaction between User, Frontend, and Backend, the BERT module, Hybrid Decision Engine, and Database during cyberbullying detection.

1.User Input Submission: It starts with the user inputting text into the system via the frontend interface. This text represents the content that needs to be analyzed for cyberbullying.

2.FrontendTransmission of Requests: The frontend side validates input data and sends this data to the backend through an API request for the consistency of data in transit.

3.Initiating the backend processing: Upon receiving the request, the backend prepares the input for analysis and forwards the text to generate contextual embedding in the BERT module.

4.Embedding Response: The computed embeddings are returned from the BERT module to the backend for classification processing.

5.Classification Request to Hybrid Engine: The backend sends these embeddings to the hybrid detection module for classification. The classification is performed using a combination of rules and BERT predictions.

6.Prediction Generation: Finally, the hybrid produces a final label (for instance, bullying or non-bullying), along with a confidence score, which is then passed on to the backend.

7.Result Logging: The result, confidence, and other related data are stored in the database to maintain auditing and further analysis. Response Preparation: The backend will then have a response with a classification label. Result Display: The response is received by the frontend, that shows the user the detection result.

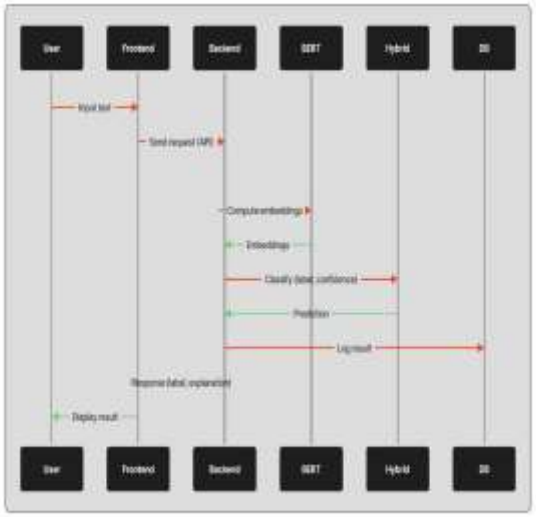

Figure 3 : Sequence flow

VI. CORE FUNCTION & MODULE

The DetectNet system consists of various defined modules that are responsible for different tasks in the overall task of detecting cyberbullying. The basic functions of the DetectNet model and their respective modules have been defined as follows:

1) User Interface Module: This module creates an interactive form that allows the end user or moderator to input textual data for analysis. This includes input validation, error management, and display of classification outcomes with explanations.

2)Frontend Communication Module: The frontend handles client-side processing and ensures data transmission in a secure manner to the backend using RESTful APIs. It provides consistency in data formatting and ensures timely data transmission, i.e., low latency.

3)Backend Processing Module: The backend acts as the central controller of the system. It coordinates tasks such as preprocessing, invocation of BERT model and rule-based engine, data flow between these modules, and so on.

4)Text Preprocessing Module:This module carries out various text normalization, tokenization, stop-word filtering, emoji interpretation, and noise reduction steps to clean the text before further analysis.

5)BERT Embedding and Classification Module Description: This module uses context embeddings provided by a pre-trained BERT network, along with semantic classification techniques for detecting implicit cyberbullying messages.

6) Rule-Based Detection Module:The rule-based module utilizes predetermined lexical dictionaries, syntactic patterns, and heuristic rules for precise detection of explicit abuse, profanity, and hate speech.

7) Hybrid Decision Fusion Module: This module uses the fusion of predictions from the BERT classifier model and the rule-based detector using confidence-based fusion.

8) Database and Logging Module: The database module is responsible for holding inputs, prediction results, confidence, and timestamp information for auditing, evaluation, and improvement of the system.

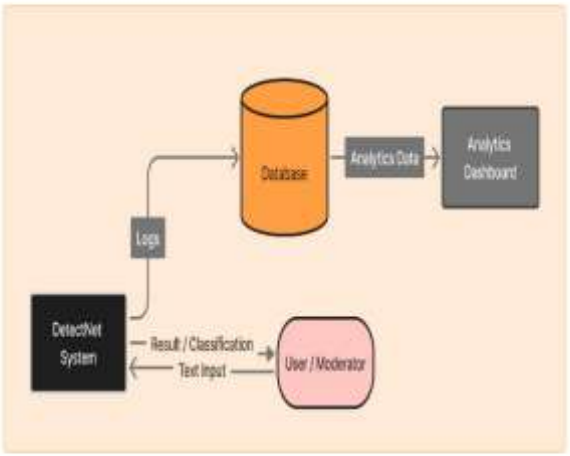

Figure 4 : DFD -0

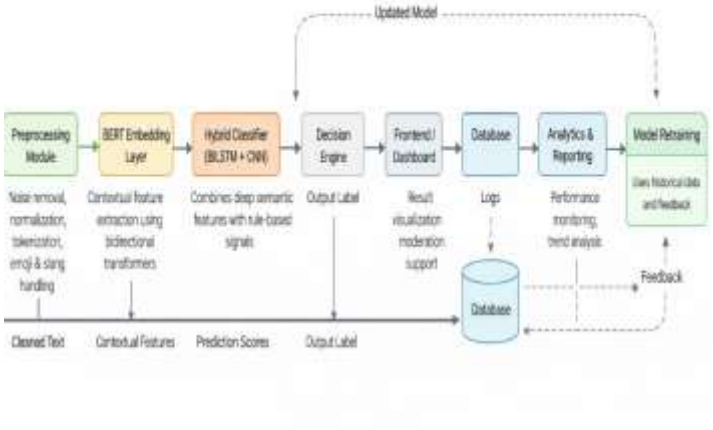

Figure 5 : DFD-1

VII. RESULTS & CONCLUSION

The reliable and consistent performance of the advocated DetectNet, which combines two approaches for cyberbullying detection, rule-based filtering, and BERT-based contextual modeling, becomes evident upon examining the comprehensive results of various evaluation parameters. It becomes clear that the rule-based approach successfully recognizes explicit abusive words and known forms of cyberbullying, which in turn leads to high precision. At the same time, the BERT model recognizes contextual dependencies and hence indirect and sarcastic cyberbullying cases.

The experimental findings show that DetectNet is better than purely rule-based systems and deep learning models with respect to its accuracy, precision, recall, and F1 score. The performance of the hybrid approach is effective and minimizes the occurrence of false positive and negative results, which is due to the inflexibility of the keywords and contextual ambiguities, respectively. Moreover, the proposed approach is effective even when working with noisy and informal text-based social media streams.

VII. FUTURE SCOPE

Though DetectNet has proved its effectiveness for detecting cases of cyberbullying, there are many avenues for improvement which can be considered in order to further enhance its viability for real-world scenarios. In the future, work can be done to enhance

the DetectNet system for detecting language variants such as mixed language, especially regional languages, which are often used in social media today. Incorporating features such as multimodal analysis can also help enhance the detecting capabilities of the system to a larger extent.

Furthermore, if strategies such as adaptive learning and model updating processes are included, it could increase its effectiveness against evolving slang, new forms of abusive patterns, and concept drift. Additionally, the inclusion of explainable AI (XAI) could be utilized to increase transparency and trust by providing users with comprehensible insights into the model decisions made.
Future enhancements may also extend to the deployment of the DetectNet in a real-time and large-scale environment using distributed and edge computing. Lastly, the employment of fairness-aware learning and mitigation techniques may ensure that the detection of cyberbullying by the DetectNet model is ethical and responsible across various user groups, hence making it a socially impactful solution.

VII.REFERENCES

1. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., Polosukhin I. “Attention Is All You Need”, 2017 https://arxiv.org/abs/1706.03762

2. Devlin J., Chang M. W., Lee K., Toutanova K. — “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, 2019 https://arxiv.org/abs/1810.04805 arXiv+1

3. Rosa H., Pereira N., Ribeiro R., Ferreira P. C., Carvalho J. P., Oliveira S., Coheur L., Paulino P., Veiga Simão A. M., Trancoso I. — “Automatic Cyberbullying Detection: A Systematic Review” — 2019 — https://doi.org/10.1016/j.chb.2018.12.021 Universidade Lusófona Research1

4.Almerekhi.H, Ahmed.M & Farouq I. "BERT-ox: BERT-Based Toxic Comment Classification" 2024

5. Singh R , & Verma A. "Detecting Harassment in Indian Languages Using Transformers" 2024

6. Sharma P, Kulkarni S, & Das R, "CyberTox: Benchmarking LLMs for Online Abuse Detection" 2024

7. Fortuna P, & Nunes S, "A Survey on Automatic Detection of Hate Speech in Text" 2018

8. Rosa.H, Pereira.N, Ribeiro R, Ferreira P, CarvalhoJ. P, OliveiraS, SilvaT, & Coheur, L. "Automatic Cyberbullying Detection: A Systematic Review" 2019

9. Hosseinmardi H, Mattson, S. A Rafiq, R. I. Han, & Mishra, S. "Detection of Cyberbullying Incidents on the Instagram Social Network" 2015

10. YinD, Xue Z, Hong L, Davison B. D, Kontostathis A, & Edwards L "Detection of Harassment on Web 2.0" 2009 Thangaraj, S., Karthick, R., & Kumar, **V.,** “AI-Driven Chatbot for Justice: Bridging the Legal Accessibility Gap,” IEEE Access, vol. 12, pp.