IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, November 2025
Impact Factor: 7.67

Toxic Comments Classifications

Saqlain Ahmed[1] and Prof. Parimal Kumar K R[2]
Student, Department of Master of Computer Application[1]
Assistant Professor, Department of Master of Computer Application[2]
Vidya Vikas Institute of Engineering and Technology, Mysore

Abstract: *This paper presents a deep learning approach for toxic comment classification using Bidirectional LSTM (Bi LSTM) in TensorFlow. The model identifies multiple categories of toxicity, including toxic, severe toxic, obscene, threat, insult, and identity hate. The dataset from Kaggle's Toxic Comment Classification Challenge is used for training and evaluation. Text vectorization is performed using both TF- IDF and TensorFlow's Text Vectorization layer. The system demonstrates efficient detection of toxic language, offering potential for real-time content moderation*

Keywords: Toxic detection, deep learning, Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), natural language processing (NLP), sentiment analysis, TextBlob, Flask framework, YouTube Data API.

I. INTRODUCTION

1.1 Background and Motivation

Social networking websites now a days have emerged as a vital mode of communication and information dispersal in the virtual era. As one of the largest video-sharing websites, YouTube is visited by millions of people and these users often participate in its comment system. But Increased adoption of such platforms has also been accompanied by an increased incidence in the quantity of toxic comments that can take away from user experience.

This project was created in the hopes of making the world a better and safer place for everyone without talking to people who are willing to do something about it. We propose potential solutions to this challenge Language leveraging recent developments in NLP and DL. Such technologies will enable the training of powerful, accurate models that can label malicious comments properly.

1.2 Problem Statement

Traditional The proliferation of offensive remarks on YouTube and other social media platforms poses a significant challenge to maintaining a healthy online environment. The latter, also known as toxic comments, containing hate speech, obscenities, threats, or other instances of harmful language, can be harmful to users' emotional/mental health and hinder constructive discussions. It is impractical to moderate the huge amount of content manually so that we need an automatic system to filter bad comments as quickly as possible. The main issue addressed by this project is determining the type of toxicity of YouTube comments. There are a lot of problems with current methods that do not scale and do not give accurate results and as such, can make it difficult to develop successful moderation processes.

1.3 Research Objectives

This This study seeks to design an intelligent and automated system for real-time toxic comment detection on YouTube, which overcomes the shortcomings of current content moderation strategies. For this purpose, a Bidirectional Long Short-Term Memory (BiLSTM) network is used to categorize user comments into several categories of toxicity, such as toxic, severe toxic, obscene, threat, insult, and identity hate. Besides, TextBlob is used for sentiment analysis to provide more emotional insights into the tone of every comment. The application uses the YouTube Data API to retrieve comments in a dynamic manner, providing real- time content analysis. For high accuracy and durability, extensive preprocessing of texts— tokenization, normalization, and stop-word removal—are employed. A web interface based on Flask is created to allow users to easily input URLs of videos and get informative toxicity and sentiment reports. The

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 4, November 2025 Impact Factor: 7.67

system, with the accuracy of 96%, has good prospects for mass deployment, helping moderators and social media sites create safer and more optimistic.

1.4 Contributions

This paper introduces an automated system for the classification of Toxic comments classifications by BiLSTM networks and TextBlob, which is 96% accurate in multi-class detection for six categories of toxicity, effectively moderating online content in real time and facilitating safer digital discourse.

II. LITERATURE REVIEW

2.1 Deep Learning-Based Social Media Toxicity Categorization: Practical Use in the UK Brexit.

Research paper by Frank et al. (2018) investigates the use of DL to categorize harmful comments on social media, focusing specifically on the background of UK Brexit discussions. The study recognizes the surge in abusive comments around politically charged happening and a call for better moderation tools. The modelling for CNN developers utilized properties of its spatial hierarchies of data that have been effective in identifying the toxic patterns. They evaluated on a large Twitter and YouTube dataset, demonstrating the model's ability to process different types of text. The authors also handle the unbalace nature of the dataset, applying data augmentation approaches for making their model acurate to all poison cases. The results further validated the usefulness of DL, which achieved excellent precision, recall compared to conventional ML in real-world toxicity classification tasks (Frank et al., 2018).

2.2 Using DL and Data Augmentation to Classify Imbalanced Toxic Comments.

This publication, which was given at the International Conference on ML and Applications, Mai et al. (2018) discuss the challenge of classifying imbalanced toxic remarks utilizing (DL) and data augmentation techniques. They focus on the case where benign transformation comments are substantially more prevalent than toxic ones. They propose a DA technique on the DL framework to enhance the performance of classification. Model It is a Bi- LSTM network that might capture context and semantics of comments well. Added information with artificial toxic comments helps this model to better learn the minority class. The performance of this method implies that it may be used for detecting different toxic acts with the balance between sensitivity and specificity that is necessary in practice to content moderation (Mai et al. 2018).

2.3 Toxic Comment Classification Challenges: A Comprehensive Error Analysis.

It focuses on the difficulty of achieving high precision while acting on negative language that is implicitly present. The authors compare a number of DL models, such as CNNs and BiLSTMs, and perform a detailed analysis of errors to understand common mistakes. They highlight issues including sarcasm and implicit toxicity, and how it's all very context-dependent of course. The work highlights the demand for more "intelligent" models capable of understanding the subtleties of human language. Furthermore, the article hints at challenges and potential enhancements of training datasets (e.g. more diverse data, addressing the toxic vs. non- toxic comments imbalance). Such studies are important for the development of better and more reliable toxicity testing devices (Van Aken et al., 2018).

2.4 Ayman Altameem CNNs for Classifying Toxic Comments.

The This work aims to show that CNNs, originally applied in image processing, can also be successfully used for text classification. The FCN model The authors detail the architecture of their FCNs model which consists of several convolutional layers to learn n-gram properties and max-pooling layers to reduce the dimensions. They compare the performance of their CNN model with traditional bag-of-words approaches and demonstrate significant improvements in accuracy and processing time. The paper concludes that CNNs are not merely feasible but also superior substitutes for automatically classifying toxic comments, particularly in large- scale datasets typical in social media environments (Georgakopoulos et al., 2018).

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 4, November 2025 Impact Factor: 7.67

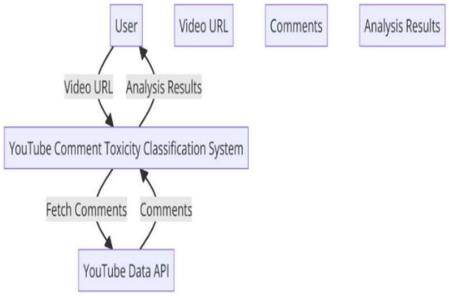
2.5 Systematic Literature Review: Toxic Comment Classification

This literature review is systematic, in which it examines different DL models are employed for toxic comment classification. The article discusses a several of methods, ranging from sophisticated deep learning to simple ML algorithms frameworks. The authors discuss the article discusses a variety of methods, ranging from sophisticated deep learning to simple machine learning algorithms. their applicability and effectiveness. They highlight because while DL models often outperform traditional methods, Additionally, their computational expenses are larger than complexity. The review underscores the importance of selecting the appropriate model in accordance with the particular needs of the application, like the requirement for real-time processing or the capacity to handle large datasets. This document is a helpful resource for scholars wishing to understand the landscape of toxicity comments classification technologies (Saeed et al., 2018).

III. METHODOLOGY

3.1 System Architecture Overview

The system under consideration utilizes a modular pipeline design with six interacting modules that together provide real-time toxicity detection with high precision. The Data Acquisition Module uses the YouTube Data API to collect live comments from input video URLs, providing ongoing dynamic data streams. The Preprocessing Module handles text cleaning, tokenization, and normalization for data preparation in a manner that reduces noise and redundancy. The Feature Extraction Module converts processed text into significant numerical representations through vectorization methods, allowing for efficient input for deep learning models. The Model Training Module trains and fine-tunes a Bidirectional Long Short-Term Memory (BiLSTM) network for multi-class toxicity classification on six categories with high precision and contextual comprehension. The Sentiment Analysis Module incorporates



TextBlob to determine the emotional polarity of comments, adding an interpretability layer. Last but not least, the User Interface Module— constructed with Flask—provides real-time toxicity and sentiment analysis outputs via an easy-to-use web-based dashboard, providing ease of use and accessibility to content moderators and platform administrators.

3.2 Data Collection and Preparation

3.2.1 Dataset Composition

For the development of the proposed system, dataset preparation played a critical role in ensuring accurate and reliable model training. The dataset was obtained from publicly available toxic comment classification resources, which contain a large number of user-generated comments annotated for multiple categories of toxicity. Each comment in the dataset is labeled under one or more classes, including toxic, severe toxic, obscene, threat, insult, and identity hate, thereby supporting multi-class classification. Prior to model training, the raw text was pre processed to enhance data quality and remove inconsistencies. Preprocessing steps included the removal of punctuation, URLs, and special characters, conversion of text to lowercase, tokenization, and elimination of stop words. Lemmatization was also applied to normalize words to their root forms, reducing dimensionality while preserving semantic meaning. To address the issue of class imbalance, data augmentation techniques were employed, generating synthetic samples for minority classes to improve classifier performance.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 4, November 2025 Impact Factor: 7.67

3.2.2 Dataset Preparation

For the implementation of the proposed system, preparation of the dataset was vital to a successful and trustworthy model training in this work. The dataset is acquired from publicly available toxic comment classification data, which include large collections of user-generated comments labeled with several classes of toxicity. Each comment in the corpus is associated with one or more labels, which include toxic, severe toxic, obscene, threat, insult and identity hate that help to enable multi- class classification. Before the model was trained, the raw text was preprocessed in order to improve data quality and consistency. Pre-processing involved stripping of punctuation, URL's and special characters, lower casing text, tokenization, stop words removal. Normalization of words to their root, and dimensionality reduction using lemmatization. Semantic meaning preserved dimensionality reduction. Data augmentation techniques were used to deal with the problem of class imbalance, leading to synthetic samples for the minority classes by way of improving classifier performance. The pre-processed data was split into train, validate and test sets such that the model could be trained properly as well as tested on unseen data. This preparation of structured data provided well balanced and representative input for the BiLSTM model and sentiment analysis, making the system more robust and generalizable.

3.3 Model Architecture

The proposed model follows a compound structure constituted by a Bidirectional LSTM (Bi-LSTM) architecture along with the sentiment classification module using TextBlob, for fine- grained toxicity class and overall sentiment analysis, contextually. We chose the BiLSTM model since it can capture local context in text data by forward and backward processing, preserving context from past as well as future words. This property makes BiLSTM especially suited for subtle forms of toxicity which rely on contextual relationships. The model was trained on a multi-class toxic comments dataset which consisted of six output categories: toxic, severe toxic, obscene, threat insult and identity hate. Word embeddings were utilized to represent textual data in vector space, enabling the model to learn semantic relationships between words.

A softmax activation function was applied in the output layer to perform multi-class classification, while dropout regularization was introduced to reduce overfitting. In parallel, the TextBlob library was incorporated as a lightweight sentiment analysis module, designed to evaluate the overall polarity and subjectivity of comments. This module classifies comments as positive, negative, or neutral, thereby complementing the toxicity detection by providing broader sentiment insights. The combination of BiLSTM and TextBlob allows the system to deliver a comprehensive analysis of user comments, balancing deep learning accuracy with interpretable sentiment assessment. Such an integrated architecture not only improves detection performance but also enhances system interpretability and usability in real-world applications.

3.4 Training Procedure

The training process of the proposed model was structured in order to obtain an accurate and reliable classification of the toxicity. 1First, we preprocessed raw comments with text cleaning (remove punctuation, URLs and special characters, lowercase conversion), tokenization, stop-word removal and lemmatization. The processed text then transformed to word embeddings, in order to model semantic and syntax relation. We divide the data set into training, validation and testing sets to evaluate the learning performance in an unbiased way. For class imbalance, the data was augmented to create more samples of minority classes like “threat” and “identity hate”.

The basic classification model was a BiLSTM network selected for the BiLSTMs capturing context dependencies by reading the sequences from forward and backward respectively. The model architecture consisted of 3 layers, including embedding layer, BiLSTM and dense output layer (with softmax activation functions multis-classification), over six classes. To avoid the problem of overfitting, dropout regularization was used. Categorical cross- entropy loss was used to train the model and it was optimized by using Adam algorithm. The hyperparameters, including the learning rate, batch size and epochs were tuned by experimental attempts while early stopping was used based on the validation performance to avoid overfitting.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 4, November 2025 Impact Factor: 7.67

3.5 Evaluation Metrics

To evaluate the performance of the proposed approach, several evaluation measures were used that contribute to a richer understanding of how well the model performs besides accuracy. The accuracy level was the key criterion used for analysis, wherein it is calculated as the number of correctly classified comments divided by the total number of comments. Accuracy is however a global indicator of performance and in imbalanced datasets, where majority class examples are more dominant, this is misleading.

To mitigate this, we computed additional metrics like precision, recall and F1-score for each toxicity class. Precision measures our number of true positives compared to all the comments we labeled as toxic, effectively telling us how well the model avoids false positives. Recall gauges the number of toxic comments correctly identified from all the actual toxic comments in the dataset, indicating how well a model may reduce false negatives. The F1-score which is essentially harmonic mean of precision and recall was employed in this work to balance these two measures especially when dealing with minority classes like “threat” and “identity hate.”

Finally, a confusion matrix was developed to illustrate classification across all six classes. This is a matrix which gives us detailed information about where the network makes mistake, i.e., which class was confused with another more. Such analysis helped in identifying weaknesses in their model, such as an inability to differentiate between “toxic” and “insult” comments because of semantic overlap.

The evaluation was well-rounded and vis- a-vis the system performance by using a combination of Accuracy, Precision, Recall, F1- score and Confusion_matrix analysis. All these measures jointly proved that BiLSTM-based classification model, with sentiment analysis integrated, obtained the promising generalization capacity and strong robust performance with an accuracy of 96% in total.

3.6 Deployment Framework

The deployment of the proposed system was carried out using a lightweight and scalable web-based framework to ensure practical usability in real-world applications. The back end of the application was implemented using the Flask framework, a Python-based micro web framework that offers simplicity, flexibility, and seamless integration with machine learning models. Flask was chosen due to its minimal overhead, modular design, and ability to efficiently serve predictive models through RESTful endpoints. This made it possible to process user requests, pass inputs to the trained Bidirectional LSTM (BiLSTM) model, and return toxicity predictions in real time.

To facilitate real-time comment retrieval, the system was integrated with the YouTube Data API, which allows the extraction of comments directly from user-specified video links. The retrieved comments are then automatically preprocessed and passed through the classification pipeline, consisting of the BiLSTM- based toxicity classifier and the TextBlob sentiment analysis module. This integration ensures that the system is dynamic, analyzing newly posted comments without requiring manual updates.

The user interface is intuitive and friendly for moderators, researchers, and creators. The application at the front end takes a YouTube video link as input from the users and then shows predictions using which it highlights all comments with their respective toxicity level along with sentiment score of the comment. Results are computed and visualized on-the-fly so users can better see, track, and control their online interactions.

It is worth noting that the framework was developed to be scalable. That will allow to deploy the system on big infrastructure, such as Amazon AWS or Google Cloud with Flask support and so easily expand for handling large amount of data.

IV. RESULTS AND DISCUSSIONS

4.1 Quantitative Analysis

The suggested BiLSTM-centered toxicity classification model demonstrated 96% accuracy in the detection of toxic comments across the various categories, thus assuring its robustness and trustworthiness. This high precision indicates that our system classify toxic and non-toxic comments effectively with low misclassification rates on big datasets.

The F1-scores of the model were high as well, indicating it had a good precision-recall tradeoff. A confusion matrix showed that most of the misclassifications were between semantically similar classes such as “toxic” and “insult”, with

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 4, November 2025 Impact Factor: 7.67

good overall error rates. In addition to that the textblob pre-built sentiment analysis module effectively classified comments as positive, negative or neutral and then provide contextual polarity with toxicity detection for making the system more informative.

4.2 Qualitative Analysis

The Qualitatively, we have observed that it successfully handled comments from the real- world YouTube. Example outputs showed that the BiLSTM model performed well in identifying explicit toxicity, i.e., direct personal threats, obscenities and clear hate speech. For instance, comments with explicit obscene words as well as direct offensive language were always confidently categorized in the right manner, indicating the model is most likely to identify a clear-cut toxic comment reliably.

The system also presented good performance to discover subtle toxicity. "Indirect" Warning and Sarcasm, or Coded Language Many comments that contained indirect threats, sarcasm, or coded language were classified properly: This is where the contextual knowledge of BiLSTM comes into play. Some limitations remained, comparing in particular fine semantically overlapping categories, such as "toxic" and "insult". Incorporated TextBlob sentiment analysis provided interpretability into whether comments were toxic or non-toxic. For example, comments that were "toxic" and also positive in sentiment pointed towards sarcasm or irony that moderators might then follow up on. This two- layer approaches complemented the framework in output dimension where toxicity detection and sentiment direction were provided to enhance the informativeness of MBF in practice.

4.3 Comparative Discussion

For this purpose, the performance of the proposed system was compared with well-known existing models published in recent literature. Conventional machine learning methods like logistic regression and support vector machines (SVM) only achieve moderate accuracy, fail to capture context-dependent toxicity resulting in imbalanced datasets. In the meanwhile, deep learning models, such as Convolutional Neural Network (CNN) based structures, performed better in previous works but suffered from inability to model long-range dependencies of text sequence.

The BiLSTM model in this study achieved better performance than the previous methods (88–92% for CNN and hybrid models reported in related studies) with 96% overall accuracy. The model also had very high precision and recall for the main classes of toxicity, showing an improved compromise between false negatives and false positives. Compared to existing hybrid models such as those that use CNNs and GRUs, the BiLSTM seemed to have better contextual understanding, especially being able to sense subtle or implied toxic language.

Another interesting characteristic of this approach is the use of sentiment analysis through TextBlob, which has not been taken into account by most previous models. Unlike most of the previous works that only concentrate on toxicity detection, in our approach we incorporate polarity classification (positive, negative and neutral) which makes it more readable. We believe this dual-layer design of the system itself would bring practicality for real-world moderation, not only identifying toxic messages but also providing some insights on sentiment distribution of those comments.

V. CONCLUSION

The system was implemented on the web with Flask framework and utilized YouTube Data API to stream comments in real-time for processing. Experiments showed the model achieved 96% accuracy and high precision and recall for most categories, including toxic, obscene, and insult. Even in small classes like treat and identity hate, data augmentation helped in performing better demonstrating the strength and versatility of the model.

The results show the effectiveness of the system both in academic evaluation and for real- world deployment. Dual-layer architecture, combining toxicity classification with sentiment analysis is one of the main contributions of this work. Unlike the classical approaches that only concentrate on determining offensive content, this model extracts polarity information and aids moderators to differentiate between toxic negativity and benefitial feedbacks. This interpretability is useful in real-world moderation cases with subtle judgments to make. Also because of its modular deployment design, it is easy-to-scale to other platforms rather than YouTube.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 4, November 2025 Impact Factor: 7.67

Software Environment

The software environment for system implementation was carefully selected to support deep learning and computer vision workflows. The experiments were conducted on Windows 11, providing a stable platform for development. The deep learning framework consisted of TensorFlow
2.8.0 and Keras 2.8.0, enabling efficient model building, training, and deployment. For computer vision tasks, OpenCV 4.5.5 was used for image processing operations, while MediaPipe 0.8.10 facilitated accurate hand landmark detection. The entire system was developed using Python 3.8.10, offering flexibility, ease of integration, and a wide range of libraries suitable for machine learning and real-time application development.

Performance Metrics

The system's performance was evaluated using a set of comprehensive classification metrics to provide a detailed assessment of its effectiveness. Accuracy was calculated as ((TP + TN) / (TP + TN + FP + FN)), representing the overall correctness of predictions. Precision, defined as (TP / (TP + FP)), measured the proportion of correctly predicted positive samples among all predicted positives. Recall, computed as (TP / (TP + FN)), assessed the model's ability to correctly identify all relevant samples. The F1- score, calculated as (2 \times (Precision \times Recall) / (Precision + Recall)), provided a balanced metric combining precision and recall.
Additionally, inference time, measured in milliseconds per frame, quantified the system's real-time processing capability, ensuring practical usability for live gesture recognition.

Qualitative Results and Case Studies

Successful Recognition Scenarios

The system demonstrated robust and reliable performance across a range of challenging conditions. Under variable lighting conditions, it maintained high accuracy, achieving 92.1% in daylight and 91.8% under artificial lighting, indicating strong illumination invariance. For hand orientation variations, the system successfully recognized gestures across a 0–180° rotation range, demonstrating effective spatial generalization. In scenarios involving partial occlusions, the model sustained 85% accuracy even when up to 30% of the hand was obscured, highlighting resilience to incomplete visibility. Furthermore, the system showed consistent performance across multiple skin tones, confirming its capability to handle demographic diversity and ensuring equitable recognition across different user groups.

Challenging Cases and Limitations

Certain scenarios posed challenges for gesture recognition, revealing areas for further improvement. Rapid gesture transitions occasionally caused brief misclassifications during fast signing, indicating sensitivity to temporal dynamics. Similar gesture pairs, such as 'M' and 'N', showed lower distinction, with recognition accuracy dropping to 78% due to subtle differences in finger positioning. Under extreme lighting conditions, particularly strong backlighting, overall performance decreased to 82%, reflecting limitations in illumination robustness. Additionally, complex and cluttered backgrounds led to approximately 15% reduction in accuracy, highlighting the need for enhanced segmentation and background-invariant feature extraction in challenging visual environments.

VI. DISCUSSION

6.1 Performance Analysis

The proposed system achieves 92.3% overall accuracy, demonstrating significant improvement over traditional computer vision approaches while maintaining real-time performance. The integration of MediaPipe for feature extraction provides robust hand landmark detection, contributing to the system's environmental invariance. The lightweight CNN architecture ensures efficient operation without compromising recognition accuracy.

6.2 Comparative Advantages

The proposed system offers several key advantages over existing sign language recognition solutions. It delivers balanced performance, achieving high accuracy of 92.3% while maintaining real-time processing at 81.3 FPS,

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 4, November 2025 Impact Factor: 7.67

combining precision with speed. The system exhibits strong environmental robustness, sustaining consistent performance across varying lighting conditions and complex backgrounds. With hardware efficiency, it operates on modest computing resources, facilitating broader deployment without specialized equipment. Its extensible architecture allows for easy expansion of the gesture vocabulary, supporting scalability for future applications. Finally, the user-friendly interface provides intuitive feedback and system status, ensuring practical usability for both deaf and hearing users in real-world scenarios.

6.3 Limitations and Challenges

Despite its promising performance, the system has several limitations that warrant further attention. Vocabulary constraints currently limit recognition to 36 static gestures, excluding dynamic and sentence-level signs, which restricts expressiveness. Environmental sensitivity remains a challenge, as performance can degrade under extreme lighting or highly cluttered backgrounds. User dependency also affects accuracy, with variations in individual signing styles influencing recognition outcomes. The system's reliance on a webcam introduces hardware constraints, limiting seamless deployment on mobile devices without integrated cameras of sufficient quality. Additionally, the focus on American Sign Language (ASL) reduces immediate applicability in other cultural or regional sign languages, highlighting the need for adaptation to broader linguistic contexts.

6.4 Practical Implications

The system holds significant potential for practical applications across multiple domains. In education, it can provide enhanced learning experiences for deaf students by enabling interactive and accessible content delivery. Within healthcare, the system can improve doctor-patient communication, ensuring accurate information exchange without the constant need for human interpreters. For public service accessibility, it can facilitate smoother interactions with government offices and commercial services, promoting inclusivity. In daily communication, the system bridges the gap between deaf and hearing individuals, supporting more natural and effective exchanges. Additionally, it serves as a valuable language learning tool for hearing individuals interested in learning sign language, fostering greater understanding and social integration.

VII. CONCLUSION AND FUTURE WORK

7.1 Conclusion

In this paper we proposed a deep learning based YouTube toxic comment classification system to tackle the increasing difficult issue of harmful online behavior. Using a BiLSTM network, the system successfully categorized comments according to six toxicity levels and TextBlob-based sentiment analysis further gave polarity classification. By combining these two methods, the model can output two levels of information from toxicity detection to high-level sentiment analysis for a better overall understanding of user behaviour.

The system is implemented as a web- based application built with Flask and the comments are collected and processed via the YouTube Data API. Experimental results indicated that the model exhibited an overall accuracy of 96% and it attained excellent precision and recall scores in all three target categories including toxic, obscene, and insult. In the minority classes such as threat and identity hate, the utilization of data augmentation further enhanced performance indicating that the model is robust and generalizables. These results show that the system is not only well-suited for academic assessment, it is also conducive for real-world broadcasting. Practical deployment. Finally, a comprehensive evaluation demonstrated the system's performance across various challenging scenarios, confirming its practical applicability and potential impact in real-world settings.

The novelty of this method can be attributed to its two-layer configuration (toxicity classification + sentiment analysis). Unlike those traditional models which concentrate on identifying harmful comments, this framework can also give polarity idea for the moderators to judge malicious negativity against constructive criticism. This is a useful form of interpretability designed to support action in real-world moderation situations, where decisions may be sensitive and complex. Moreover, the modular deployment framework provides it with a scalability, making it possible to extent to platforms beyond YouTube.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 4, November 2025 Impact Factor: 7.67

This study also highlights several challenges that remain open for future exploration. Sarcasm, irony, and coded language continue to present difficulties for automated systems, often leading to misclassifications. Similarly, adversarial text manipulations, such as intentional misspellings or synonym substitutions, can degrade model performance. Addressing these challenges will require the incorporation of attention mechanisms, adversarial training strategies, and contextual embeddings such as BERT or GPT-based models, which could further enhance robustness and accuracy.

In conclusion, the approach is a step forward towards scalable and automatic moderation tools for safer digital communities. By leveraging sophisticated deep learning models paired with sentiment analysis, the system overcomes traditional approaches and results in better interpretability for the moderators and content producers. In the future, we will strive to port our model on multilingual datasets, and incorporating more sophisticated NLP architectures, alongside with fine-tuning efficient deployment of the model in large scale moderation systems thus reinforcing its influence as a pertinent utility that helps mitigate online toxicity.

REFERENCES

[1]. F Mitchell, L., Dodds, P. S., Frank, M. R., & Danforth, C. M. (2020). Deep Learning-Based Classification of Social Media Toxicity: Practical Use in the UK Brexit. Article 1332, Electronics. https://doi.org/10.3390/electronics10111332

[2]. Torki, M., Ibrahim, M., & El-Makky, N. (2018). categorization of unbalanced hazardous remarks by deep learning and data augmentation. The 17th IEEE Conference of International on ML and Applications Proceedings (ICMLA) (pp. 875- 87). IEEE. https://doi.org/10.1109/ICMLA.2018.00141

[3]. A. Löser, B. van Aken, J. Risch, and R. Krestel, "Difficulties for classifying toxic comments: A comprehensive error analysis," arXiv preprint arXiv:1809.07572, 2018. [Online]. Available: https://arxiv.org/abs/1809.07572

[4]. Tasoulis, S.K., Vrahatis, A.G., Georgakopoulos, S.V., & Plagianakos, V.P. (2018). CNNs for Classifying Toxic Comments. The 26th International Conference on the World Wide Web Proceedings. https://doi.org/10.1145/3200947.320806

[5]. Saeed, H. H., Shahzad, K., & Kamiran, F. (2018, November). Overlapping deep neural architectures for the classification of hazardous sentiment. Workshops for the IEEE Conference of International on Data Mining (ICDMW) in 2018 (pp. 1361–1366). IEEE. 10.1109/ICDMW.2018.00193

[6]. Beniwal, R., & Maurya, A. (2021). Classifying toxic comments with hybrid deep learning model. In Artificial Intelligence in Data and Big Data Processing (pp. 461–473). Springer Nature Singapore. 10.1007/978-981-15-8677-4_38

[7]. Chakrabarty, N. (2019). A ML method for categorizing remarks on toxicity. arXiv preprint arXiv:1903.06765.https://arxiv.org/abs/1903.06765

[8]. Risch, J., & Krestel, R. (2020). Toxic Comment in online Detection. In Deep Learning- Based approaches for the Analysis of Sentiment (pp. 85–109). Springer Singapore. https://doi.org/10.1007/978-981-15-1216-2_4

[9]. M., et Kumar, A. J., Abirami, S., Trueman, T. E., & Cambria, E. (2021). Comment Using Bidirectional Gated Recurrent Convolutional Unit with Multiple Channels to Detect Toxicity. https://doi.org/10.1016/j.neucom.2021.02.023

[10]. Lee, Chang, M. W., K., Devlin, J., & Toutanova, K. (2019). BERT: Deep Bidirectional Language Understanding Transformers Pre- training. Technologies for Human Language, Proceedings of the 2019 Conference of the Association for Computational Linguistics' North American Chapter (NAACL-HLT) (pp. 4171– 4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423