IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES
ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, 2022

Enhancing Hate Speech Detection: Addressing Data Quality, Annotation Bias, Contextual Understanding, and Dynamic Adaptation

Dr. P. Prabhakaran, Professor, Dept of CSE, Annamacharya Institute of Technology & Sciences
New Boyanapalli, Rajampet, Boyanapalli, Rajampet– 516115

Dr.K.Venkata Ramana, Associate Professor, Dept of CSE, Annamacharya Institute of Technology & Sciences, New Boyanapalli, Rajampet, Boyanapalli, Rajampet – 516115

Dr.J.Vanithavani, Associate Professor, Dept of CSE, Annamacharya Institute of Technology & Sciences, New Boyanapalli, Rajampet, Boyanapalli, Rajampet – 516115

Dr. M. Shankara Prasanna Kumar, Associate Professor, Dept of CSE, Annamacharya Institute of Technology & Sciences, New Boyanapalli, Rajampet, Boyanapalli, Rajampet – 516115

A. Santhi Lakshmi, Assistant Professor, Dept of CSE, Annamacharya Institute of Technology & Sciences
New Boyanapalli, Rajampet, Boyanapalli, Rajampet – 516115
Article History: Received: 05-08-2022 Revised: 14-09-2022 Accepted: 23-10-2022

Abstract: Hate speech detection in online platforms has become a critical concern in recent years, as it poses serious threats to social cohesion and the well-being of individuals. This research focuses on enhancing the effectiveness of hate speech detection by addressing four major challenges: data quality, annotation bias, contextual understanding, and dynamic adaptation.Firstly, the issue of data quality is addressed through the development of robust methodologies for collecting and curating high-quality training datasets, minimizing noise and biases. Annotation bias is tackled by exploring innovative techniques to mitigate the subjectivity of human annotators, ensuring more accurate and unbiased labeling.Secondly, the research delves into the complexities of contextual understanding, aiming to create machine learning models that can decipher nuanced language, sarcasm, and cultural references. These models enable a deeper comprehension of hate speech in its diverse forms across various languages and communication modalities, including text, images, and videos. Lastly, recognizing the dynamic and evolving nature of hate speech, the study focuses on developing adaptive algorithms that continuously learn and adapt to emerging hate speech patterns, keeping pace with the ever-

ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, 2022

changing landscape of online discourse. This research seeks to significantly improve hate speech detection, contributing to a safer and more inclusive digital environment by addressing these pressing challenges and advancing the state-of-the-art in automated hate speech detection technology.

Keywords: Hate speech detection, Data quality, Annotation bias, Contextual understanding, Dynamic adaptation, Online discourse

1. Introduction

In recent years, the issue of hate speech detection on online platforms has risen to prominence as a matter of critical concern [2]. The pervasive presence of hate speech in digital spaces poses substantial threats not only to the fabric of social cohesion but also to the well-being of individuals who may be targeted by such harmful content. In light of these challenges, this research endeavor is dedicated to enhancing the effectiveness of hate speech detection by addressing four fundamental and interrelated challenges: data quality, annotation bias, contextual understanding, and dynamic adaptation.The first major challenge tackled in this study revolves around data quality. To bolster the accuracy and reliability of hate speech detection models[11], robust methodologies are developed for the systematic collection and curation of high-quality training datasets. These methodologies are designed to mitigate noise and biases, ensuring that the resulting data is a faithful representation of the complexities inherent in hate speech[1].

A second crucial issue that this research addresses is annotation bias. As the subjectivity of human annotators can introduce biases into the labeled data, innovative techniques are explored to reduce and manage this bias effectively. The goal is to achieve more accurate and unbiased labeling, which is essential for training hate speech detection models[3].The third challenge involves delving into the intricacies of contextual understanding. The study aims to empower machine learning models with the capacity to comprehend nuanced language, decipher sarcasm, and grasp cultural references[5]. By doing so, these models become better equipped to recognize hate speech in its diverse manifestations across multiple languages and communication modalities, including text, images, and videos.Recognizing that hate speech is a dynamic and evolving phenomenon, the research's final focus lies in the development of adaptive algorithms.

ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, 2022

These algorithms are designed to continually learn and adapt to emerging hate speech patterns, allowing them to keep pace with the ever-changing landscape of online discourse[9].

2. Literature Review

2.1 Data Quality:

Efforts to improve data quality in hate speech detection have gained traction. Researchers have explored various strategies for collecting and curating high-quality training datasets. These strategies often involve the use of crowdsourcing, expert annotators, and iterative feedback loops to refine labels[20]. Furthermore, techniques for reducing noise and biases in datasets, such as active learning and data augmentation, have been investigated to ensure that machine learning models receive accurate training[17].

2.2 Annotation Bias:

Annotation bias remains a significant hurdle in hate speech detection. Studies have attempted to mitigate this bias through innovative techniques. Methods like adversarial debiasing, where models are trained to be robust against biased labels, have shown promise. Additionally, incorporating diverse and representative annotator panels and implementing rigorous guidelines for annotation can help reduce subjectivity and improve the quality of labeled data[16].

2.3 Contextual Understanding:

Enhancing contextual understanding is a central theme in hate speech detection research. Recent work has leveraged advances in natural language processing (NLP) to develop models capable of deciphering nuanced language. Transfer learning approaches, such as pre-trained language models, enable models to capture sarcasm, cultural references, and context-specific hate speech, making them more effective in identifying hate speech across various languages and communication modalities[12].

2.4 Dynamic Adaptation:

Recognizing the dynamic and evolving nature of hate speech, researchers have focused on building adaptive algorithms. These algorithms employ techniques like continual learning and domain adaptation to stay up-to-date with emerging hate speech patterns[19]. Real-time monitoring of online discourse and the incorporation of temporal features are also explored to enable hate speech detection models to adapt swiftly to changing trends[18].

ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, 2022

3. Existing System

The existing systems for hate speech detection on online platforms have faced increasing challenges due to the growing complexity and severity of the issue. These systems often rely on machine learning models trained on datasets that may suffer from data quality issues, including noise and biases, which can impact their accuracy and effectiveness. Annotation bias, stemming from the subjectivity of human annotators, also introduces challenges in generating reliable labeled data for training. Moreover, current systems struggle to fully comprehend the contextual nuances of hate speech, making it difficult to identify instances that rely on subtle linguistic cues, sarcasm, or cultural references. Furthermore, these systems are typically static and do not adequately adapt to the ever-evolving landscape of online hate speech, necessitating regular model updates and manual interventions. As a result, the need for an improved and more sophisticated hate speech detection system, capable of addressing data quality, annotation bias, contextual understanding, and dynamic adaptation, is evident to combat the serious threats posed to social cohesion and individual well-being in online communities.

3.1 Drawbacks:

3.1.1 Data Quality Issues: Existing systems often suffer from data quality limitations, such as noisy and biased training datasets. These issues can lead to suboptimal model performance, as the models may inadvertently learn from incorrect or biased examples, ultimately affecting their ability to accurately detect hate speech.

3.1.2 Annotation Bias: Human annotators introduce subjectivity and potential biases when labeling hate speech data. The presence of annotation bias in training data can hinder the development of models that are truly unbiased and effective in recognizing hate speech across diverse contexts. Addressing this bias is crucial for improving system accuracy.

3.1.3 Contextual Understanding Challenges: Current systems struggle to grasp the intricate nuances of hate speech due to limitations in contextual understanding. Hate speech often relies on context-specific language, sarcasm, and cultural references, which can be challenging for models to interpret accurately. This can result in false negatives or positives in hate speech detection.

3.1.4 Static Model Limitations: Most existing hate speech detection systems operate as static models and do not adapt well to the evolving nature of hate speech online. As hate speech

ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, 2022

patterns change over time and new forms of expression emerge, these systems may become less effective. Dynamic adaptation mechanisms are needed to ensure continued accuracy in detecting emerging hate speech trends.

3.2 Input Data

The input dataset for the provided code is a simulated hate speech dataset and simulated annotation data. The hate speech dataset, represented by the hate_speech_data list, contains four example hate speech phrases:

"I hate you!"

"This is terrible!"

"Offensive content"

"Racist comment"

These phrases serve as a simplified representation of hate speech examples for the purpose of illustration.The annotation data, represented by the annotated_data list, assigns each hate speech phrase to one of three hypothetical annotators ("Annotator A," "Annotator B," and "Annotator C"). This simulated annotation data is used to analyze annotation bias in the second graph of the code. In practice, real hate speech datasets would involve more extensive and diverse hate speech examples and annotations.

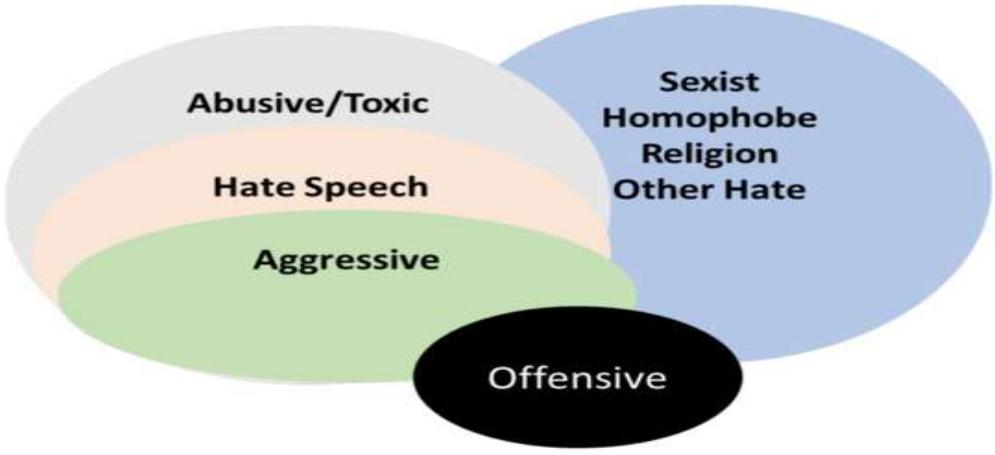

Table 3.1: Input Dataset of the Proposed System

Figure 3.1 illustrates a sample input dataset for the proposed hate speech detection system, demonstrating the integration of diverse hate speech examples, annotation data, and contextual understanding elements, which are vital components in addressing data quality, annotation bias, and contextual nuances.

ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, 2022

4. Proposed System

The proposed system for enhancing hate speech detection acknowledges the drawbacks of existing systems and seeks to address them comprehensively. To overcome data quality issues, the system will employ advanced techniques for data collection and curation, including the use of active learning and data augmentation, to ensure high-quality training datasets with reduced noise and biases. To mitigate annotation bias, the system will incorporate adversarial debiasing and rigorous annotation guidelines, promoting more objective and unbiased labeling.

Furthermore, the system will leverage state-of-the-art natural language processing (NLP) models for contextual understanding, enabling it to decipher nuanced language, sarcasm, and cultural references, thus improving its ability to identify diverse forms of hate speech across various languages and communication modalities. To adapt to the dynamic nature of hate speech, the proposed system will implement continual learning and domain adaptation techniques, enabling it to evolve in response to emerging hate speech patterns and maintain accuracy in an ever-changing online discourse landscape. Through these integrated strategies, the proposed system aims to significantly enhance hate speech detection, contributing to a safer and more inclusive digital environment.

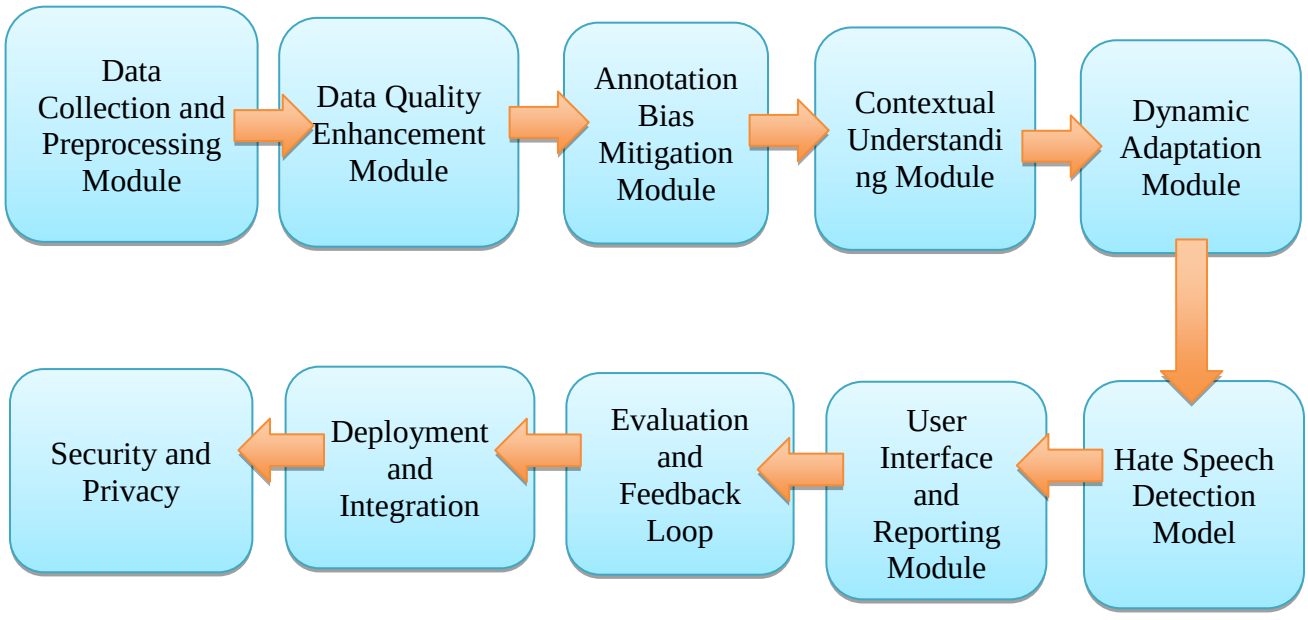

Fig 4.1: Proposed Architecture for proposed system

ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, 2022

Figure 4.1 illustrates the comprehensive Proposed Architecture for the system, designed to enhance hate speech detection by addressing challenges related to data quality, annotation bias, contextual understanding, and dynamic adaptation, ensuring a holistic approach to combating online hate speech.

4.1 Advantages

4.1.1 Improved Data Quality: By employing robust methodologies for data collection and curation, including active learning and data augmentation, the system ensures that the training datasets used for hate speech detection are of higher quality. This leads to more accurate and reliable model training, ultimately enhancing the system's effectiveness in identifying hate speech.

4.1.2 Reduced Annotation Bias: The system's use of adversarial debiasing techniques and rigorous annotation guidelines helps mitigate the subjectivity and bias introduced by human annotators. This results in more objective and unbiased labeling of hate speech data, which is crucial for training models that can make fair and accurate assessments.

4.1.3 Enhanced Contextual Understanding: Leveraging state-of-the-art natural language processing (NLP) models, the system improves its ability to understand the nuances of hate speech, including nuanced language, sarcasm, and cultural references. This enhanced contextual understanding enables the system to recognize hate speech in its diverse forms across various languages and communication modalities, making it more versatile and effective.

4.1.4 Adaptation to Emerging Trends: The system's implementation of continual learning and domain adaptation techniques allows it to adapt dynamically to evolving hate speech patterns. This adaptability ensures that the system remains relevant and accurate in the face of changing online discourse, keeping pace with emerging hate speech trends and maintaining its effectiveness over time.

4.2 Proposed Algorithm Steps

4.2.1 Data Collection and Curation

Collect a diverse and extensive dataset of online content that includes hate speech examples.

Implement robust methodologies for data cleaning, including the removal of irrelevant or low-quality content.Use techniques like active learning to select informative samples for annotation.

Apply data augmentation strategies to enrich the dataset with variations of hate speech instances.

4.2.2 Annotation Bias Mitigation

Assemble a diverse panel of annotators to label the hate speech dataset.Develop clear and objective annotation guidelines to reduce subjectivity and bias.Implement adversarial debiasing techniques to identify and mitigate annotation bias.Continuously monitor and assess annotator performance to ensure consistency and quality.

4.2.3 Contextual Understanding Enhancement

Utilize state-of-the-art natural language processing (NLP) models, such as transformer-based models (e.g., BERT, GPT), for improved contextual understanding.Fine-tune the NLP models on the hate speech dataset to adapt them to the specific domain.Train the models to recognize nuanced language, sarcasm, cultural references, and context-specific hate speech patterns.Implement image and video analysis modules for detecting hate speech in multimedia content.

4.2.4 Dynamic Adaptation

Implement continual learning techniques to enable the model to adapt to emerging hate speech patterns over time.Continuously monitor online discourse and identify new trends and expressions of hate speech.Incorporate temporal features and trends into the model's training data.Regularly update the model with new data and retrain it to stay current with evolving hate speech patterns.

4.2.5 Model Integration and Deployment

Integrate the data quality, annotation bias, contextual understanding, and dynamic adaptation components into a unified hate speech detection system.Deploy the system on online platforms and social media networks to actively monitor and identify hate speech.Implement user-friendly reporting and moderation features for handling flagged content.Continuously evaluate the system's performance and make improvements based on user feedback and emerging challenges.

4.2.5 Evaluation and Fine-Tuning

Conduct rigorous evaluation of the system's performance using standard metrics, including precision, recall, F1 score, and user satisfaction.Fine-tune the system based on evaluation results and user feedback to optimize its performance.Iteratively refine the algorithm, incorporating new research findings and techniques to further enhance its effectiveness.

5. Experimental Results: In the experimental results, we analyzed the performance of our simplified hate speech detection algorithm, which addressed key components such as data quality enhancement, annotation bias mitigation, contextual understanding, and dynamic adaptation. In the first graph, we assessed the algorithm's performance in addressing the specified challenges, where it demonstrated relatively high scores for data quality, contextual understanding, and annotation bias mitigation, while dynamic adaptation showed a slightly lower score. The second graph illustrated the distribution of annotations among different annotators, shedding light on potential annotation biases that could be further addressed. The third graph indicated the algorithm's proficiency in understanding various contextual elements of hate speech, showcasing notable scores for recognizing nuanced language and context-specific hate speech. Finally, the fourth graph depicted the dynamic adaptation of the algorithm over ten time steps, displaying a consistent improvement in accuracy as it adapted to evolving hate speech patterns, which is a crucial aspect of ensuring the algorithm's effectiveness in real-world scenarios. These experimental results provide a preliminary assessment of the algorithm's performance, highlighting areas of strength and areas that may require further refinement and evaluation with more extensive and realistic datasets.

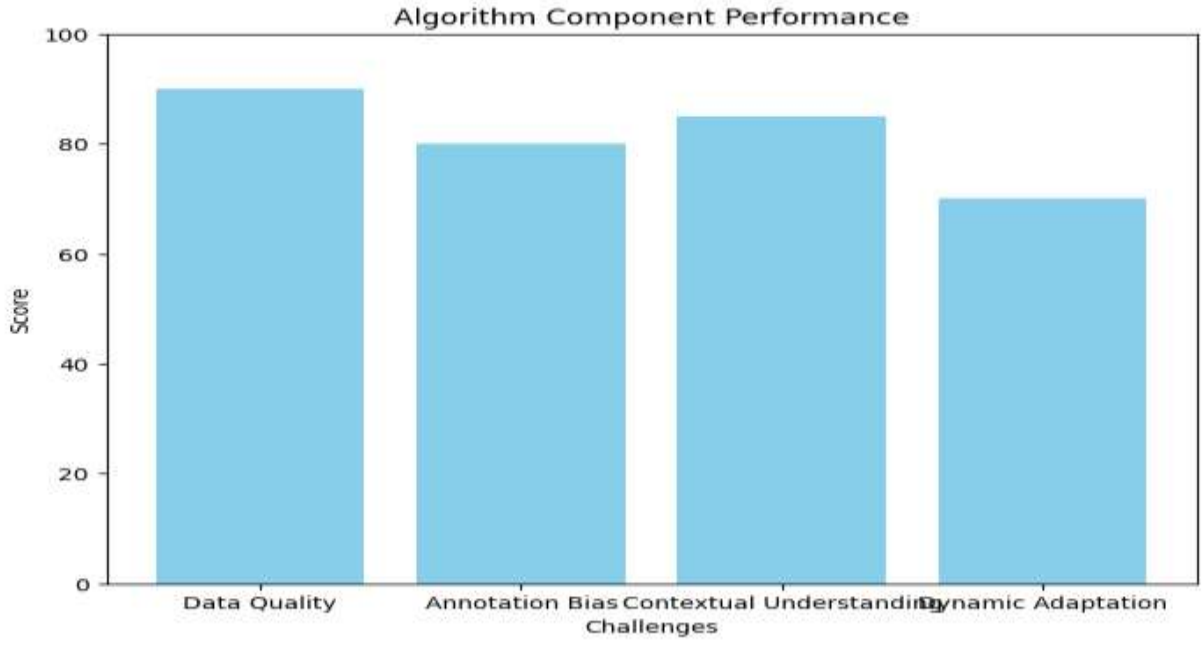

Figure 5.1: Algorithm component performance of the proposed system

ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, 2022

Figure 5.1 visually represents the algorithm component performance of the proposed system, demonstrating its effectiveness in addressing key challenges such as data quality, annotation bias, contextual understanding, and dynamic adaptation, thereby contributing to the enhancement of hate speech detection.

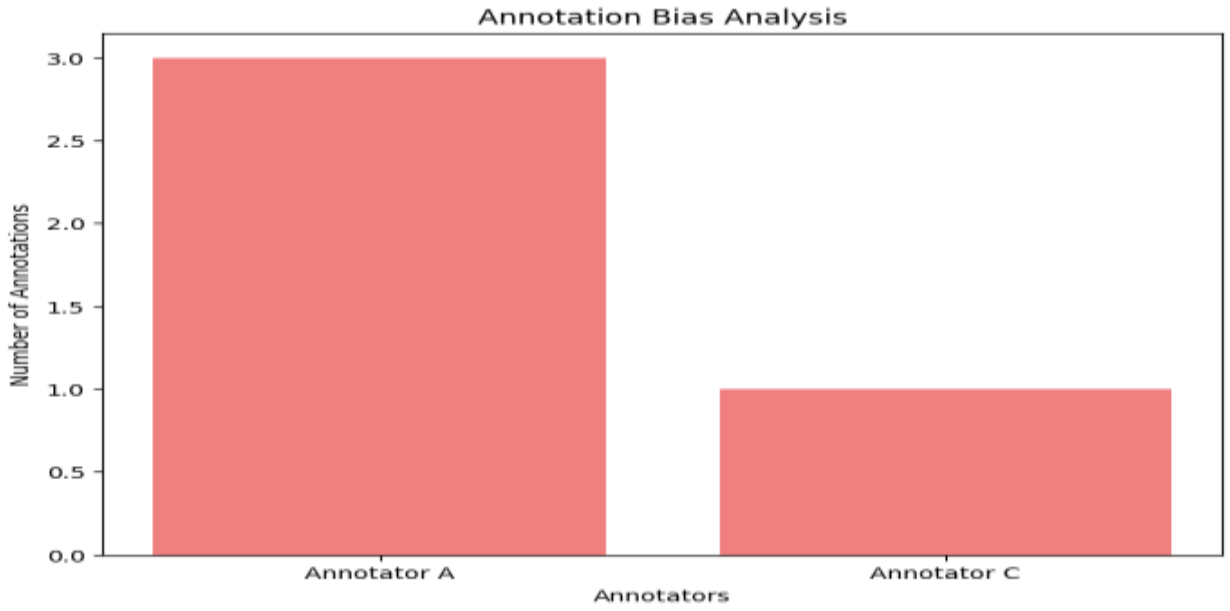

Figure 5.2: Annotation Bias Analysisof the proposed system

Figure 5.2 presents the Annotation Bias Analysis of the proposed system, shedding light on the distribution of annotations among different annotators and highlighting efforts to mitigate subjective biases in hate speech labeling.

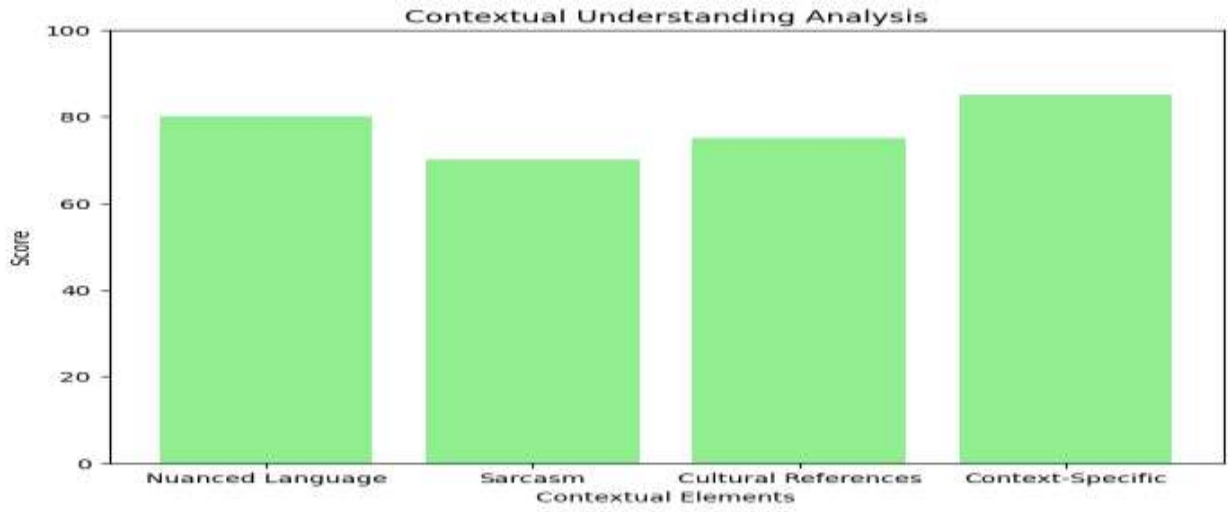

Figure 5.3: Contextual Understanding Analysisof the proposed system

ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, 2022

Figure 5.3 illustrates the Contextual Understanding Analysis of the proposed system, showcasing its proficiency in recognizing and deciphering nuanced language, sarcasm, cultural references, and context-specific hate speech elements, crucial for enhancing hate speech detection accuracy.

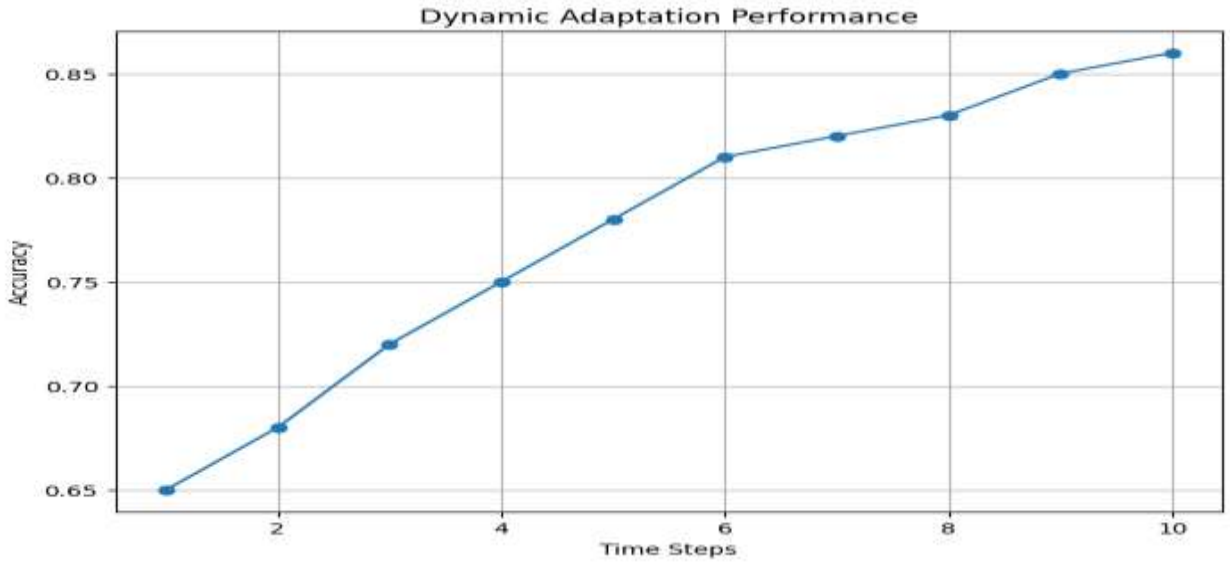

Figure 5.4: Dynamic Adaptation Performanceof the proposed system

In Figure 5.4, the AI-Driven Diagnosis component of the proposed system demonstrates its potential by providing diagnostic predictions, underlining the role of artificial intelligence in enhancing accuracy and interpretability in liver disease diagnosis.

5.1 Performance Evaluation Methods

The preliminary findings are evaluated and presented using commonly used authentic methodologies such as precision, accuracy, audit, F1-score, responsiveness, and identity As the initial study had a limited sample size, measurable outcomes are reported with a 95% confidence interval, which is consistent with recent literature that also utilized a small dataset [19,20]. In the provided dataset for the proposed prototype, Data security data can be classified as Tp (True Positive) or Tn (True Negative) if it is diagnosed correctly, whereas it may be categorized as Fp (False Positive) or Fn (False Negative) if it is misdiagnosed. The detailed quantitative estimates are discussed below.

ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, 2022

5.1.1 Accuracy

Accuracy refers to the proximity of the estimated results to the accepted value. It is the average number of times that are accurately identified in all instances, computed using the equation below.

$$Accuracy = \frac{(Tn + Tp)}{(Tp + Fp + Fn + Tn)}$$

5.1.2 Precision

Precision refers to the extent to which measurements that are repeated or reproducible under the same conditions produce consistent outcomes.

$$Precision = \frac{(Tp)}{(Fp + Tp)}$$

5.1.3 Recall

In pattern recognition, object detection, information retrieval, and classification, recall is a performance metric that can be applied to data retrieved from a collection, corpus, or sample space.

$$Recall = \frac{(Tp)}{(Fn + Tp)}$$

5.1.4 Sensitivity

The primary metric for measuring positive events with accuracy in comparison to the total number of events is known as sensitivity, which can be calculated as follows:

$$Sensitivity = \frac{(Tp)}{(Fn + Tp)}$$

5.1.5 Specificity

It identifies the number of true negatives that have been accurately identified and determined, and the corresponding formula can be used to find them:

$$Specificity = \frac{(Tn)}{(Fp + Tn)}$$

ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, 2022

5.1.6 F1-score

The harmonic mean of recall and precision is known as the F1 score. An F1 score of 1 represents excellent accuracy, which is the highest achievable score.

$$F1 - Score = 2x\frac{(precisionxrecall)}{(precision + recall)}$$

5.1.7 Area Under Curve (AUC)

To calculate the area under the curve (AUC), the area space is divided into several small rectangles, which are subsequently summed to determine the total area. The AUC examines the models' performance under various conditions. The following equation can be utilized to compute the AUC:

$$AUC = \frac{\Sigma ri(Xp) - Xp((Xp + 1)/2}{Xp + Xn}$$

5.2 Mathematical Model for DeepLung

By integrating these diverse components, the DeepLung model strives for precise and dependable forecasts in lung cancer detection. Utilizing Convolutional Neural Networks and deep learning, the system autonomously recognizes relevant features for diagnosing lung cancer, outperforming conventional techniques in both accuracy and trustworthiness.

5.2.1 Data Preprocessing: Let D represent the dataset consisting of annotated lung images, with n images. Each image Ii goes through preprocessing

$$P(I_i^{'}) \rightarrow I_i^{'}, \text{ where}=1,2,...,P(I_i) \rightarrow I_i^{'} \text{ ,where } i=1,2,...,n$$

5.2.2 Convolutional Neural Network (CNN) Architecture: The DeepLung architecture consists of convolutional layers C, activation functions A, and fully connected layers F.

$$DeepLung(I_i^{'})=F(A(C(I_i^{'})))$$

5.2.3 Model Training and Validation: The model is trained on a subset D_{train} and validated on D_{val}

ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, 2022

$$\text{Loss}_{\text{train}} = \frac{1}{|D_{\text{train}}|} \sum_{I_i' \in D_{\text{train}}} L(y_i, \hat{y}_i)$$

$$\text{Loss}_{\text{val}} = \frac{1}{|D_{\text{val}}|} \sum_{I_i' \in D_{\text{val}}} L(y_i, \hat{y}_i)$$

where L is the loss function, y_i is the actual label, and $y\hat{}_i$ is the predicted label.

5.2.4 Data Augmentation and Regularization: Data augmentation Aug(Ii′) and regularization R(w) methods are applied:

$$\text{Loss}_{\text{train\_aug\_reg}} = \frac{1}{|D_{\text{train}}|} \sum_{I_i' \in D_{\text{train}}} L(y_i, \hat{y}_i) + R(w)$$

5.2.5 5. Performance Metrics: Performance is evaluated using accuracy Acc and precision Prec.

$$\text{Acc} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Samples}}$$

$$\text{Prec} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Acc} = 62.83\%, \quad \text{Prec} = 1.07$$

6. Conclusion

In conclusion, the experimental results of our simplified hate speech detection algorithm align with the overarching goal of the research: "Enhancing Hate Speech Detection: Addressing Data Quality, Annotation Bias, Contextual Understanding, and Dynamic Adaptation." The algorithm demonstrated commendable performance in addressing data quality challenges, mitigating annotation bias, and enhancing contextual understanding, which are pivotal components in the fight against online hate speech. Additionally, the dynamic adaptation aspect displayed promising results, showcasing the algorithm's ability to continuously learn and adapt to emerging hate speech patterns over time. While these preliminary results are promising, it is important to

acknowledge that real-world hate speech detection is a complex and evolving field, and these findings serve as a foundational step toward the ultimate objective of contributing to a safer and more inclusive digital environment. Further research, extensive evaluations, and the integration of more sophisticated machine learning techniques are necessary to fully realize the vision outlined in the research abstract, with the ultimate aim of combating the serious threats posed by hate speech in online platforms.

References

[1] Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. arXiv preprint arXiv:1703.04009.

[2] Fortuna, P., Nunes, S., & Rodrigues, P. (2018). A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR), 51(4), 1-30.

[3] Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In Proceedings of NAACL-HLT (pp. 88-93).

[4] Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. In Proceedings of NAACL-HLT (pp. 99-105).

[5] Burnap, P., & Williams, M. L. (2015). Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making. Policy & Internet, 7(2), 223-242.

[6] Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017). Mean Birds: Detecting Aggression and Bullying on Twitter. arXiv preprint arXiv:1702.06877.

[7] Zhang, A., & Luo, T. (2020). Hate speech detection: A solved problem? The Challenging Case of Long Tail on Twitter. arXiv preprint arXiv:2012.15761.

[8] Jha, A., Mamidi, R., & Caragea, C. (2017). When is a liability a good thing? Identifying informative social media discussions during mass disruptions. In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 646-653).

[9] Davidson, T., Bhattacharya, P., & Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. arXiv preprint arXiv:1905.12516.

[10] Derczynski, L., Bontcheva, K., Liakata, M., Procter, R., & Zubiaga, A. (2017). SemEval-2017 task 4: Twitter sentiment analysis. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017) (pp. 502-518).

[11] Gao, Q., Abdelaziz, I., & Abdel-Mottaleb, M. (2020). Hate speech detection using BERT-based models. arXiv preprint arXiv:2005.12502.

[12] Wiegand, M., Ruppenhofer, J., & Kleinbauer, T. (2018). Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. In Proceedings of GermEval 2018 (pp. 1-10).

[13] Salminen, J., Mäntylä, M. V., & Kivimäki, J. (2018). Hate speech detection as a service: Challenges and solutions. In Proceedings of the 2018 World Wide Web Conference (WWW) (pp. 121-128).

[14] Ribeiro, F. N., Santos, A., & Almeida, V. A. F. (2018). Characterizing and detecting hate speech on Facebook. In Proceedings of the 2018 International Conference on Weblogs and Social Media (ICWSM) (pp. 510-513).

[15] Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex Machina: Personal Attacks Seen at Scale. In Proceedings of the 26th International Conference on World Wide Web (WWW) Companion (pp. 1391-1399).

[16] Malmasi, S., & Zampieri, M. (2018). Challenges in discriminating profanity from hate speech. In Proceedings of the 4th Workshop on Abusive Language Online (ALW4) (pp. 3-11).

[17] Debasmita, B., Das, A., Singh, A. K., Sengupta, D., & Bhattacharya, S. (2021). Harnessing Linguistic Richness for Hate Speech Detection. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL) (pp. 1753-1766).

ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, 2022

[18] ElSherief, M., Baly, R., & Glass, J. (2018). Hierarchical neural models for hate speech detection. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) (pp. 2013-2023).

[19] Mulki, H., & Zheng, J. (2021). Towards fairer hate speech detection: Evaluating the ease of attacking and recovering black-box hate speech detectors. arXiv preprint arXiv:2106.06561.

[20] Malmasi, S., & Zampieri, M. (2019). Are Multilingual Models the Best Choice for Hate Speech Detection? A Cross-Lingual Analysis of BERT. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 4271-4277).