International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 06 | June - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

DeepShield: A Deep Learning-Based Toxic Comment Detector

P. Sejal
2111CS020501
School of Engineering
Malla Reddy University

A.Shanthipriya
2111CS020503
School of Engineering
Malla Reddy University

Gaurav Tripathi
2111CS020529
School of Engineering
Malla Reddy University

N.Sohan Mani
2111CS020530
School of Engineering
Malla Reddy University

D. Sreeja Reddy
2111CS020540
School of Engineering
Malla Reddy University

Prof. Sivakumar
Professor, Department of AIML
School of Engineering
Malla Reddy University

1 ABSTRACT:

In the contemporary digital landscape, the prevalence of toxic comments poses a significant challenge for online platforms, impacting user experiences and fostering hostile environments. To address the issue, we introduce "DeepShield," a state-of-the-art toxic comments detector built on deep learning techniques. DeepShield aims to automatically identify and flag toxic comments within online discussions, promoting healthier online communities. Leveraging a robust deep learning architecture with attention mechanisms, DeepShield captures context and nuances in language to effectively detect toxic behavior. The system is trained on a meticulously curated dataset, encompassing a wide range of toxic language to enhance model generalization. Beyond serving as a valuable tool for content moderation, DeepShield contributes to the discourse on fostering positive online interactions. By mitigating the impact of toxic comments, DeepShield strives to create safer and more inclusive online environments for all users.The objective of DeepShield is to automatically identify and flag toxic comments within online discussions, fostering healthier online communities. The system employs a robust deep learning architecture, and attention mechanisms to capture context and nuances in language. A meticulously curated dataset is used for training, incorporating a wide range of toxic language to enhance model generalization. This toxic comments detector not only serves as a valuable tool for content moderation but also contributes to the ongoing discourse on fostering positive online interaction

2. INTRODUCTION:

2.1 Problem Definition:

In the digital realm, the rise of social media platforms has facilitated unprecedented connectivity, but it has also exposed users to toxic comments, including hate speech and harassment. This proliferation of toxic content poses a significant challenge for online communities and platforms, leading to negative user experiences and potentially harmful consequences for individuals and society as a whole. Traditional approaches to content moderation often rely on manual review by human moderators, which is time-consuming, resource-intensive, and may not scale well with the volume of user-generated content. The overarching goal of our project is to contribute to the creation of safer and more inclusive online environments by empowering platforms and communities with robust content moderation tools. By accurately identifying and flagging harmful content in real-time, our model helps mitigate the spread of online toxicity and fosters constructive dialogue among users. In doing so, we aim to not only enhance user experiences but also uphold fundamental principles of respect, civility, and diversity in online interactions.

2.2 Objective of project

The objective of the model is to leverage deep neural networks to accurately identify hate speech, toxic comments, and non-toxic comments within online discourse. By harnessing the power of deep learning techniques, the model aims to discern subtle linguistic cues indicative of harmful content, while also recognizing positive and neutral language. This approach enables the model to contribute to the creation of safer and more inclusive online environments by effectively distinguishing between different types of comments, fostering constructive dialogue while mitigating the spread of online toxicity

3. PROBLEMSTATEMENT:

In the digital realm, the rise of social media platforms has facilitated unprecedented connectivity, but it has also exposed users to toxic comments, including hate speech and harassment. This proliferation of toxic content poses a significant challenge for online communities and platforms, leading to negative user experiences and potentially harmful consequences for individuals and society as a whole. Traditional approaches to content moderation often rely on manual review by human moderators, which is time-consuming, resource-intensive, and may not scale well with the volume of user-generated content.

Moreover, the subjective nature of identifying toxic comments presents another layer of complexity for human moderators. What one person perceives as harmless banter, another may interpret as offensive or hurtful. This inconsistency in judgment can result in discrepancies in content moderation decisions and, in some cases, accusations of bias or censorship. Additionally, the sheer volume of content generated on social media platforms makes it nearly impossible for human moderators to review every post and comment effectively, leading to a backlog of unmoderated content and an increased risk of harmful material slipping through the cracks.

In response to these challenges, automated

content moderation tools like DeepShield have emerged as a promising solution. By leveraging advances in deep learning and natural language processing, these tools can analyze large volumes of text data in real-time, identifying patterns and linguistic cues associated with toxicity. Unlike human moderators, AI-powered systems like DeepShield can operate at scale, processing thousands of comments per second and flagging potentially harmful content for further review or action.

4. LITERATURE SURVEY:

1.A supervised learning based tool to identify toxic code review comments" Jaydebsaker (2022):

ToxiCR, a supervised learning based toxicity identification tool for code review interactions. ToxiCR includes a choice to select one of the 10 supervised learning algorithms, an option to select text vectorization techniques, eight preprocessing steps, and a large-scale labeled dataset of 19,651 code review comments. Two out of those eight preprocessing steps are software engineering domain specific. With our rigorous evaluation of the models with various combinations of preprocessing steps and vectorization techniques, we have identified the best combination for our dataset that boosts 95.8% accuracy and an 88.9% F1-score in identifying toxic texts. ToxiCR significantly outperforms existing toxicity detectors on our dataset.

2. CS224N: Detecting and Classifying Toxic Comments" by Kevin Khieu (2019):

This paper explores deep learning methods for online comment toxicity, including SVM, LSTM, CNN, and MLP with word and character-level embeddings. Evaluation on Kaggle's Toxic Comments dataset shows the forward LSTM model excelling in word-level binary and multi-label classification. CNN performs best for character-level classification. Overall, word-level models outperform character-level ones. Future work aims for enhanced performance using richer representations and more complex deep learning models.

3. Deep learning for religious and continent-based toxic content detection and classification" by Ahmed Abbasi.(2022):

Online platforms have amplified toxic language, prompting the need for identifying it through natural language processing. This research compares deep learning algorithms for multilabel toxic comment classification, specifically focusing on Religious and Race/Ethnicity comments. Using various word embeddings and a CNN model, the study finds CNN consistently outperforms others. The research addresses challenges like unrealistic toxicity ratings for nontoxic comments related to specific identifiers.

4. Toxic Comment Classification using Deep Learning by B. Ramesh Naidu. (2023):

Online conversation platforms enable idea exchange but also spread toxic comments. Filtering manually is impractical, necessitating models like LSTM, Character-level CNN, Word-level CNN, and Hybrid (LSTM + CNN). This study contributes to developing a web interface for classifying toxic comments, enhancing the safety of online platforms..

5. Machine learning methods for toxic comment classification" Darko Androcec. (2020):

Users generate numerous comments online, making manual moderation impractical. This study reviews 31 relevant works on toxic comment classification using machine learning. Analyzing datasets, metrics, methods, toxicity classes, and comment language, it identifies gaps and suggests future research themes for improving online toxicity detection

5. METHODOLOGY:

5.1 Existing System:

In the endeavor to manage toxic comments on online platforms, various approaches have been employed. Manual moderation, involving human moderators reviewing and removing toxic comments, is effective but time-consuming and may not scale well with the volume of content.

DeepShield promotes a healthier online environment, fostering positive interactions and user engagement. Overall, DeepShield represents a significant step forward in combating online toxicity and creating safer digital spaces for all users.

Keyword-based filters offer an automated solution by flagging or removing comments containing specific terms, but they often miss nuanced forms of toxicity. Machine learning models have emerged as a promising tool, trained on labeled datasets to classify comments as toxic or not. However, implementing and maintaining these models require significant

computational and human resources. Community reporting mechanisms empower users to report inappropriate content, which can then be reviewed by moderators or automated systems. Hybrid approaches, blending automated filtering with human moderation and community feedback, offer a comprehensive strategy for effective toxicity management, leveraging the strengths of both automated and human intervention method

5.2 Proposed System:

The proposed system, DeepShield, introduces an innovative approach to combatting toxic comments in online communities. Unlike manual moderation, which can be slow and inefficient, DeepShield leverages deep learning techniques for real-time detection of toxic language. By analyzing contextual cues and nuances in language, DeepShield can accurately identify toxic comments, even in complex or subtle forms. Trained on a diverse dataset, the system demonstrates robust generalization across various contexts and languages. This scalability makes it well-suited for handling large volumes of online content. Moreover, DeepShield's automated approach reduces the resource requirements compared to traditional machine learning models, offering a cost-effective solution for online community management. By swiftly flagging toxic comments, DeepShield promotes a healthier online environment, fostering positive interactions and user engagement. Overall, DeepShield represents a significant step forward in combating online toxicity and creating safer digital spaces for all

6 . EXPERIMENTAL RESULTS:

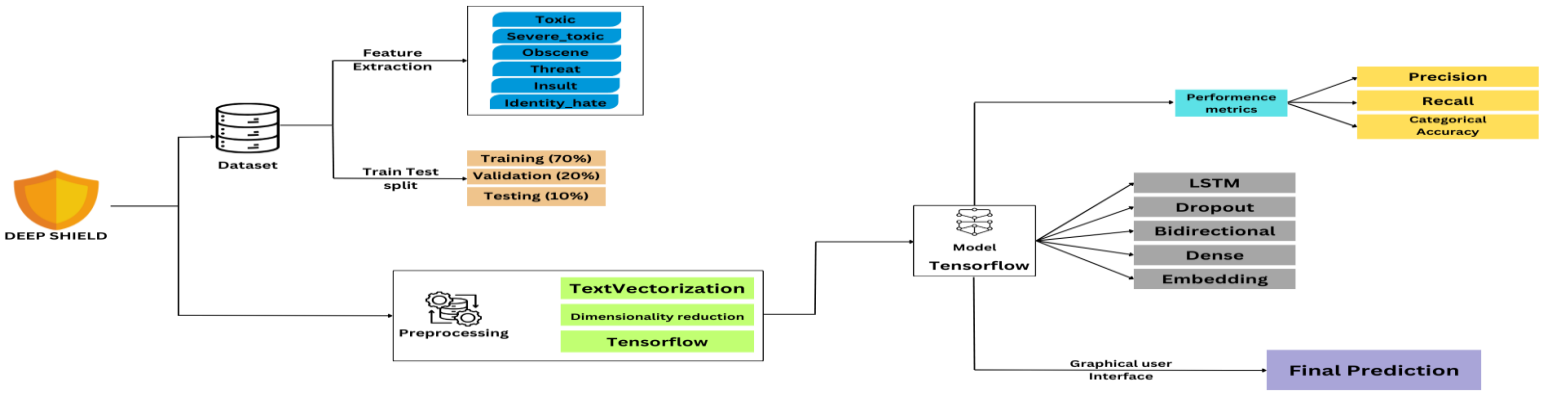

Fig 6.1 Architecture of the Deep Shield

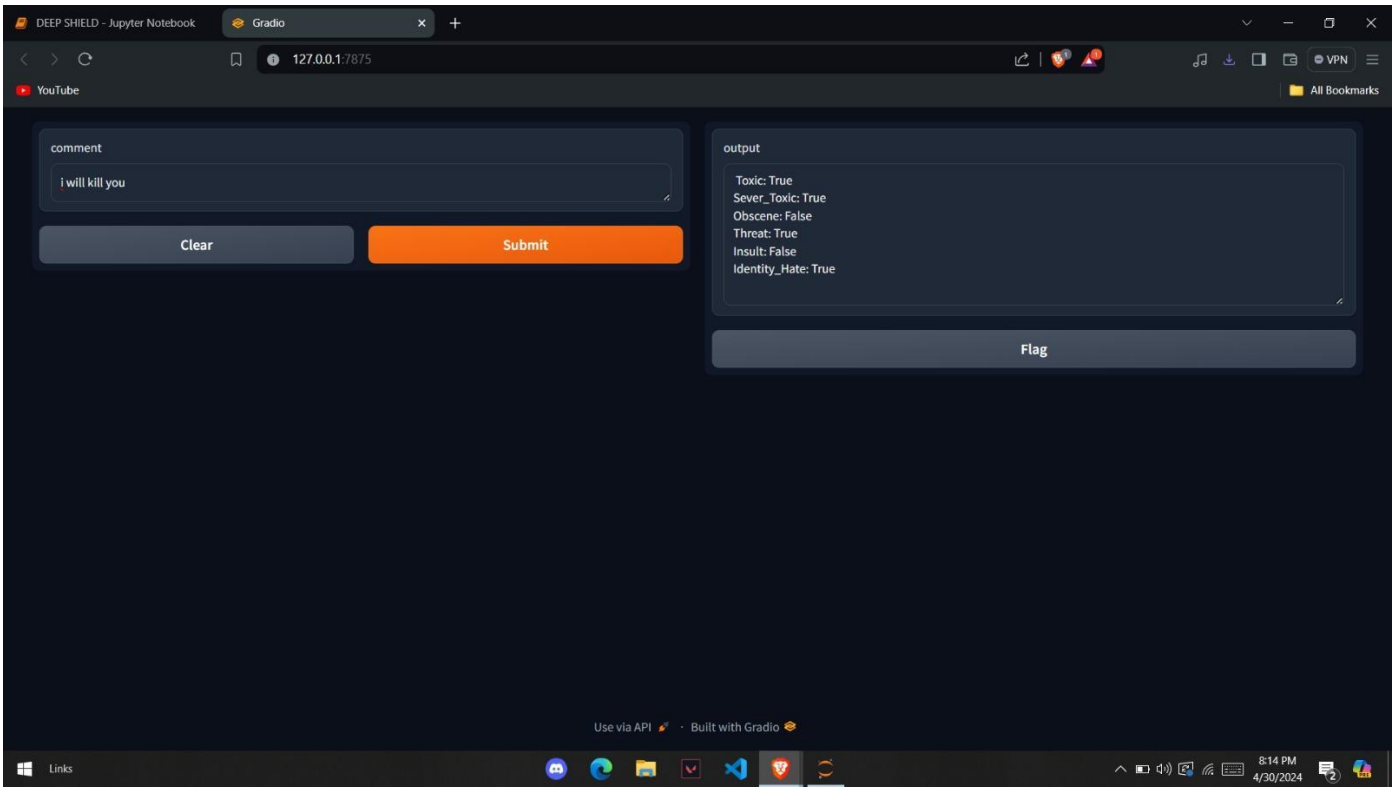

Fig 6.2 Output Screen

7 . CONCLUSION:

This project aims to automate the detection of toxic comments in online discussions. By leveraging TensorFlow/Keras, we preprocess textual data and construct a Sequential model equipped with LSTM layers for toxicity classification. The model is trained on a dataset of labeled comments, enabling it to identify various forms of toxicity, including insults, threats, and identity hate. Performance evaluation metrics such as precision, recall, and accuracy provide insights into the model's effectiveness. Once trained, the model is deployed using Gradio, allowing users to interactively test its capabilities by inputting text and receiving toxicity predictions in real-time. This project serves as a valuable tool for content moderation, empowering online platforms to maintain healthier and safer communication environments

8 . FUTURE ENHANCEMENT:

1.Attention Mechanisms: Implement more sophisticated attention mechanisms such as self-attention or multi-head attention to capture fine-grained relationships within the text data, improving the model's ability to focus on relevant information..

2.Capsule Networks: Explore the use of capsule networks, which are designed to better handle hierarchical relationships in data, for toxicity classification tasks. Capsule networks have shown promise in tasks requiring spatial hierarchies and part-whole relationships.

3.Graph Neural Networks (GNNs): Investigate the application of graph neural networks to model the relationships between words or tokens in text data. GNNs can capture structural information and dependencies, which may be beneficial for understanding the context of toxic language.

4.Multitask Learning: Explore multitask learning approaches where the model is trained on multiple related tasks simultaneously, such as sentiment analysis or hate speech detection, alongside toxicity classification. Multitask learning can help the model leverage shared representations and improve overall performance.

5.Self-Supervised Learning: Utilize self-supervised learning techniques to pre-train the model on large amounts of unlabeled data before fine-tuning on the labeled toxicity classification task. Self-supervised learning can help the model learn useful representations of the input data, leading to better generalization.

6.Dynamic Architectures: Develop dynamic architectures that can adaptively adjust their structure or parameters based on the input data or task requirements. Dynamic architectures can be more flexible and efficient, particularly in scenarios with varying text lengths or complexities.

7.Interpretability Techniques: Integrate interpretability techniques such as attention visualization or saliency maps to provide insights into the model's decision-making process.

9.REFERENCES :

1.https://github.com/conversationai/conversationai.github.io/blob/main/crowdsourcing_annotation_schemes/toxicity_with_subattributes.md

2.Kevin Khieu. Detection of harassment on web 2.0. Proceedings of the Content Analysis in the WEB 2(2019):.

3.Abbasi, A. et al. Elstream: An ensemble learning approach for concept drift detection in dynamic social big data stream learning. IEEE Access 9, 66408–66419 (2021)..

4.B. Ramesh Naidu. (2023). "Investigating Bias in Automatic Toxic Comment Detection: An Empirical Study". arXiv:2108.06487 cs.CL]