iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** V **Month of publication:** May 2025

DOI: https://doi.org/10.22214/ijraset.2025.70351

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

Offensive Language and Image Identification on Social Media Based on Text and Image Classification

G. Mahammad Idrush[1], Ch. Sai Maharsha[2], A. Gangaprasad[3], A. Sai Kuma[4]
[1]*Assistant Professor,* [2,3]*Student, Teegala Krishna Reddy Engineering College*

Abstract: ***A digital signature is like a digital version of a handwritten signature but much more secure. It ensures that digital documents are authentic, unaltered, and genuinely from the sender. Our project, Digital Signature Tool, focuses on creating an easy-to-use application for securely signing and verifying documents. Using advanced cryptographic methods like RSA or ECDSA, the tool allows users to generate and manage private and public keys securely. To sign a document, the sender uses their private key to create a unique digital signature, while the receiver uses the sender's public key to verify the signature. This process confirms the document's authenticity and ensures it has not been tampered with. The application will integrate essential features, such as secure key management, document signing, and signature verification, all within a user-friendly interface. This project aims to provide individuals and organizations with a reliable solution for protecting their documents and communications, ensuring trust, data integrity, and security in the digital space.***

I. INTRODUCTION

In recent years, social media has emerged as a primary platform for communication, expression, and information exchange. While these platforms provide a space for positive interaction and community building, they have also become breeding grounds for the spread of offensive, hateful, and abusive content. The proliferation of such harmful material—whether in the form of text, images, or a combination of both—poses serious challenges to online safety, mental well-being, and societal harmony.Detecting offensive content is a complex and evolving task due to the diversity in language, context, sarcasm, and the use of visuals. Traditional text-based moderation systems often fall short in identifying multimodal threats that combine textual and visual cues. Similarly, image-only systems may fail to understand embedded or overlaid offensive language. This has led to the growing need for robust, intelligent systems capable of analyzing and classifying both text and images to accurately identify offensive content.This research aims to develop a comprehensive framework that integrates Natural Language Processing (NLP) and Computer Vision (CV) techniques to detect offensive language and imagery on social media platforms. By leveraging advanced machine learning and deep learning models, such as transformer-based architectures for text and convolutional neural networks for images, this study seeks to improve the accuracy and reliability of offensive content detection.Furthermore, the paper explores the use of multimodal classification approaches that simultaneously process text and visual data, enhancing the system's ability to detect subtle or disguised forms of offensive content. The outcomes of this research could significantly aid social media platforms, regulatory bodies, and developers in building safer digital environments through automated moderation and content filtering systems.

II. LITERATURE REVIEW

The identification of offensive content on social media has been a critical research area in recent years, especially with the increasing use of user-generated content that includes text, images, and multimedia. Various studies have focused on tackling this issue using Natural Language Processing (NLP), Computer Vision (CV), and multimodal learning techniques.

A. Text-Based Offensive Language Detection

Early approaches to offensive language detection in text primarily relied on keyword matching and handcrafted features such as n-grams and part-of-speech tags. While these methods offered basic capabilities, they lacked context-awareness and were easily bypassed through misspellings or slang. Recent advancements in deep learning have significantly improved performance in text classification tasks. Models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), including Long Short-Term Memory (LSTM), have been used to capture contextual semantics.

However, the introduction of transformer-based models, such as BERT (Bidirectional Encoder Representations from Transformers), RoBERTa, and DistilBERT, has marked a major leap in offensive language detection. These models leverage attention mechanisms to better understand the context and semantics of language, enabling more accurate classification even in complex or implicit offensive expressions. Studies such as those by Davidson et al. (2017) introduced labeled datasets for hate speech and offensive language detection on Twitter, providing a benchmark for NLP models. Other datasets like OLID (Offensive Language Identification Dataset) from SemEval have also contributed to the development of more nuanced classification systems that consider different types of offensive behavior.

B. Image-Based Offensive Content Detection

On the visual front, offensive imagery—such as hate symbols, explicit content, or violent scenes—poses a significant moderation challenge. Traditional computer vision techniques focused on low-level features (color, texture, shape), which often failed to grasp the semantic meaning of images.

Deep learning, particularly Convolutional Neural Networks (CNNs), has transformed image classification by enabling models to learn hierarchical feature representations. Models like VGGNet, ResNet, and InceptionNet have shown strong performance in identifying offensive or inappropriate visual content. Transfer learning, wherein pretrained models are fine-tuned on smaller datasets, has been particularly effective when labeled data is limited.

Datasets like Hateful Memes Dataset (Facebook AI) and ToxiImageNet have enabled the training of image classification systems to recognize offensive imagery and visual-text combinations. These datasets highlight the importance of context—an image may appear benign but become offensive when paired with certain text.

C. Multimodal Approaches

Given the limitations of unimodal methods, recent research has shifted towards multimodal learning, which combines text and image inputs to better understand and detect offensive content. Models like VisualBERT, VilBERT, and CLIP (Contrastive Language–Image Pretraining) integrate both visual and textual representations, allowing for deeper semantic analysis.

The Hateful Memes Challenge brought attention to the challenges and importance of multimodal analysis, where hate is expressed through a combination of seemingly innocuous text and imagery. Traditional unimodal models often fail in such cases, whereas multimodal systems can learn cross-modal associations and improve detection accuracy.

Several studies have demonstrated that multimodal models outperform text-only or image-only models in detecting nuanced and context-dependent offensive content. However, challenges remain in model interpretability, data annotation, and generalizability across languages and cultures.

D. Summary and Gaps

While significant progress has been made in offensive content detection, current systems still face challenges such as:

Detecting sarcasm, irony, and slang in text.

Identifying subtle or context-dependent imagery.

Managing multilingual content and cross-cultural variation.

Ensuring real-time performance and scalability.

There is a growing need for comprehensive, multimodal systems that can robustly identify offensive content by analyzing both textual and visual inputs. This research aims to bridge these gaps by proposing an integrated framework leveraging state-of-the-art NLP and CV models for accurate and efficient offensive content identification.

III. FACTORS INFLUENCING THE SUCCESS OF PROJECT

The success of an intelligent system designed to identify offensive language and imagery on social media relies on several critical factors. These factors directly impact the accuracy, robustness, scalability, and real-world applicability of the model.

A. Quality and Diversity of Dataset

A diverse and well-labeled dataset is fundamental. The inclusion of varied linguistic styles, cultural expressions, slang, emojis, and offensive visual content ensures the model can generalize across different scenarios.

Multimodal datasets (text + image), such as the Hateful Memes Dataset, significantly improve the model's contextual understanding.

Imbalanced datasets, where offensive content is underrepresented, may lead to biased or underperforming models.

B. Effectiveness of Preprocessing Techniques

Cleaning and preprocessing text data (e.g., removing stop words, handling typos, tokenization) and image data (resizing, normalization) enhances input quality.
Techniques like lemmatization, sarcasm detection, and emoji processing further improve text classification performance.
Image augmentation helps improve generalization for visual models.

C. Selection of Appropriate Algorithms and Models

Using state-of-the-art models such as BERT for text and ResNet/VGG for images provides a strong foundation.
Multimodal models (e.g., CLIP, VisualBERT) that combine text and image embeddings are more successful in identifying context-specific offensive content.
Model architecture choice affects not just accuracy but also training time and inference speed.

D. Feature Representation and Fusion

Accurate feature extraction from both text (semantic embeddings) and images (visual features) is crucial.
The method used to fuse text and image features (early fusion, late fusion, or hybrid) influences the model's ability to make context-aware decisions.

E. Evaluation Metrics and Validation Techniques

Proper use of metrics such as Precision, Recall, F1-score, Accuracy, ROC-AUC ensures reliable evaluation.
Techniques like cross-validation and confusion matrix analysis help fine-tune the model and reduce false positives/negatives.

F. Handling Ambiguity and Context

Offensive content is often disguised through irony, metaphors, or image-text interplay. Success depends on how well the model understands and interprets implicit meaning.
Language and image context must be processed in an interconnected way to avoid misclassification.

G. Real-time Performance and Scalability

For deployment on social media platforms, the system must provide real-time detection with low latency.
The architecture should support scalability for handling large volumes of data from millions of users.

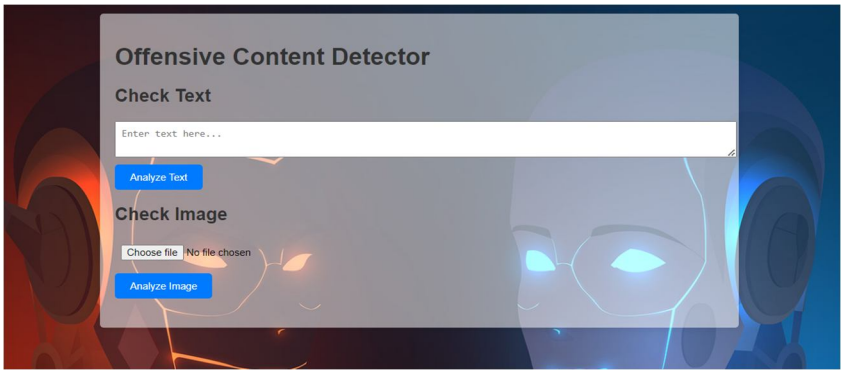
H. Ethical and Bias Considerations

Ensuring the model does not unfairly target or misclassify content based on race, gender, culture, or language is crucial.
Regular audits and use of bias mitigation techniques contribute to ethical success.

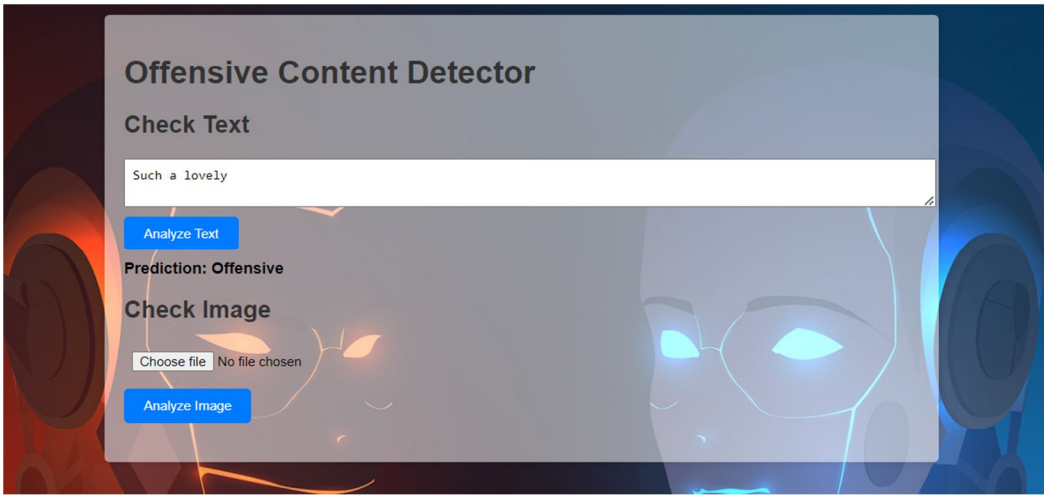
I. Continual Learning and Updates

Offensive language evolves rapidly. A successful system must support continual learning and periodic updates based on newly emerging patterns and slang

IV. OUT PUTS



Text Input Interface



Offensive Text Classification Result

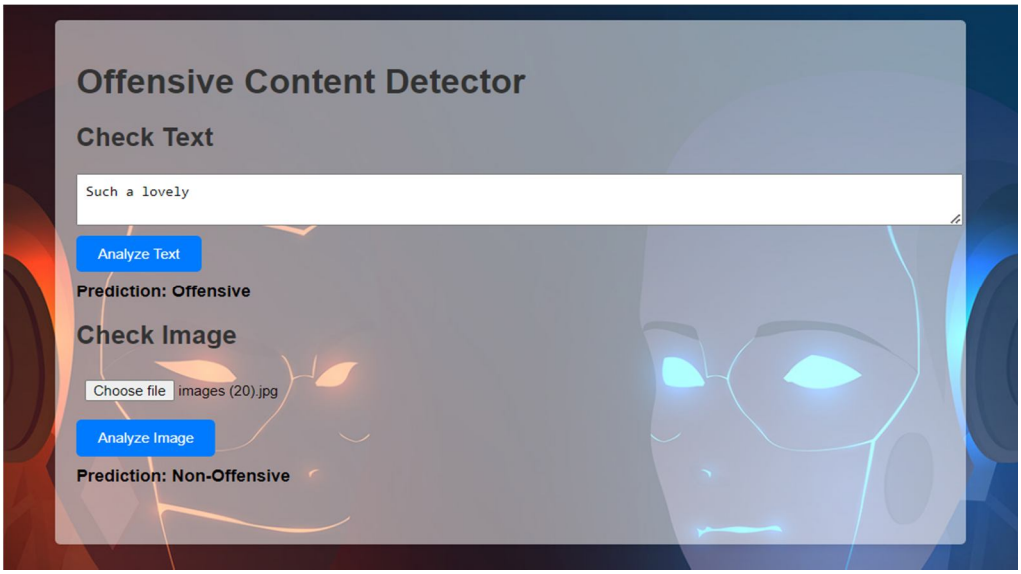


Image Classification Result

V. CONCLUSION

In today's digital age, social media platforms are increasingly becoming hotspots for the dissemination of offensive and harmful content. The proposed system leverages the power of machine learning and deep learning to address this critical issue by classifying both textual and visual content for offensiveness. By implementing separate models for text and image analysis, the system ensures a comprehensive approach to offensive content detection.

The results demonstrate that the system is capable of accurately identifying non-offensive and offensive elements in user-generated content. However, instances of false positives in text classification, such as labeling neutral sentences as offensive, indicate the need for further refinement and training on more diverse datasets. Image classification performed well, showing potential in recognizing offensive visuals, thereby enhancing the overall effectiveness of the tool.

This research highlights the feasibility of developing automated tools for real-time moderation of social media content. With further improvements in data diversity, model tuning, and the incorporation of multimodal analysis (combining text and image in a single context), the system can be a valuable asset in combating cyberbullying, hate speech, and the spread of harmful content online.

Ultimately, this project contributes to a safer digital ecosystem by empowering platforms and users with intelligent tools that promote respectful and inclusive online interactions.

VI. FUTURE ENHANCEMENTS

While the current system provides a strong foundation for detecting offensive content in both text and images, several enhancements can be introduced to improve accuracy, usability, and scalability:

1) Multimodal Content Fusion: Integrate both text and image inputs into a unified model that considers contextual relationships between text and visual content (e.g., memes). This fusion can help detect subtle and context-based offensive content.

2) Incorporation of Video Analysis: Extend the system to analyze offensive content in videos, including speech recognition for spoken language and frame-by-frame analysis for visuals.

3) Sentiment and Emotion Analysis: Enhance the text analysis module by including sentiment analysis and emotion detection to better interpret sarcasm, hidden abuse, and implied hate speech.

4) Multilingual Support: Train models to detect offensive content in regional and foreign languages, enabling the tool to cater to a broader global audience on multilingual social media platforms.

5) Adaptive Learning through User Feedback: Introduce a feedback mechanism where users or moderators can correct false positives/negatives. This input can be used to retrain and fine-tune the models periodically.

6) Real-Time Moderation API: Develop an API service for real-time integration into social media platforms, enabling automatic moderation and flagging of offensive posts during content uploads.

7) Bias Mitigation and Ethical Filtering: Ensure fairness by minimizing racial, gender, and cultural biases in training data, and regularly audit the model for ethical content moderation.

8) Explainable AI (XAI): Incorporate explainability features so that users and moderators can understand why specific content was flagged as offensive, increasing transparency and trust.

REFERENCES

[1] Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). Predicting the Type and Target of Offensive Posts in Social Media. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), pp. 1415–1420.

[2] Schmidt, A., & Wiegand, M. (2017). A Survey on Hate Speech Detection using Natural Language Processing. In Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, pp. 1–10.

[3] Basile, V., Bosco, C., Fersini, E., et al. (2019). SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 54–63.

[4] Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. In Proceedings of ICWSM 2017.

[5] Jaiswal, A., & Goel, A. (2021). Image-based Offensive Content Detection using CNN Models. International Journal of Computer Applications, Vol. 182(17), pp. 25–30.

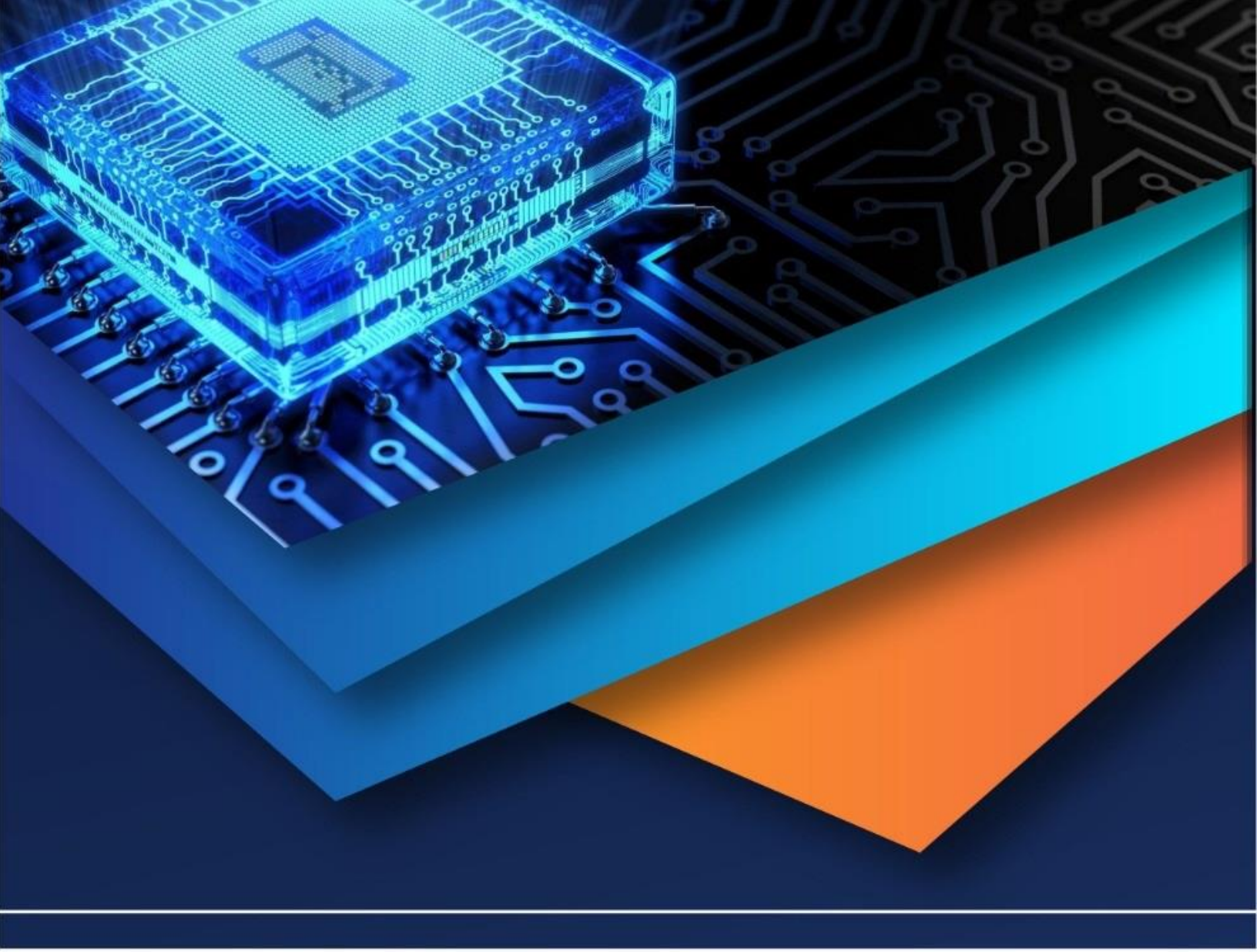
[6] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

[7] OpenAI (2020). GPT Language Models. https://openai.com/research

[8] Facebook AI (2021). Hateful Memes Challenge Dataset. https://ai.facebook.com/hatefulmemes

[9] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In Advances in Neural Information Processing Systems (NeurIPS), pp. 1097–1105.

[10] TensorFlow (2023). Open Source Machine Learning Framework. https://www.tensorflow.org



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY