Webtrace An Automated Hate Speech Detection and Enforcement System for Instagram Using Playwright, OCR, and NLP

G. Thiagarajan, Danend Krishnan U, Soniya N, Soniya N

[1]Head of the Department IT, CSI College of Engineering, Ketti, The Niligiris-643215[1]

[2]Final year IT Student, CSI College of Engineering, Ketti, The Niligiris-643215[2]

[3]Final year IT Student, CSI College of Engineering, Ketti, The Niligiris-643215[3]

[4]Final year IT Student, CSI College of Engineering, Ketti, The Niligiris-643215[4]

ABSTRACT

Social media platforms have revolutionized communication by enabling users to share their thoughts, experiences, and opinions with a global audience. Among these platforms, Instagram stands out as a widely used application that allows users to post images, videos, and stories, engage with others through comments and messages, and explore diverse content based on their interests. Its user-friendly interface and interactive features make it a preferred choice for individuals, influencers, and businesses alike. However, alongside its benefits, Instagram also presents challenges, particularly in content moderation. The open nature of the platform allows users to express themselves freely, but this can lead to the spread of harmful content, including hate speech, cyberbullying, and misinformation. To address the challenges, WebTrace introduces an automated hate speech detection framework designed specifically for Instagram. WebTrace leverages Playwright for web scraping, Optical Character Recognition (OCR) for extracting text from images, and advanced Natural Language Processing (NLP) techniques such as TF-IDF with Cosine Similarity and a USE Transformer model to analyze text content. Unlike conventional methods, WebTrace not only detects hate speech in real-time but also takes immediate action by automatically reporting the detected content to Instagram moderator.

Keywords: Hate Speech Detection, Automated Moderation, Playwright Automation, Optical Character Recognition (OCR), USE Transformer model, TF-IDF, Natural Language Processing (NLP), WebTrace.

1. INTRODUCTION

With digital technology advancing rapidly, platforms like Instagram, Twitter, and Facebook have become the main way billions of people connect, share ideas, and form communities. These platform offers discussion, creativity, and social interaction—but they also come with serious downsides. Weak content moderation policies have turned many of these spaces into hotspots for hate speech, cyberbullying, and the spread of false information. The ease of access and the ability to stay anonymous online often embolden users to post harmful or offensive content, frequently aimed at specific individuals or groups based on race, gender, religion, nationality, and other personal traits.

The impact of cyberbullying and online harassment is far from harmless. Victims may suffer from anxiety, depression, social isolation, and in some heartbreaking cases, self-harm or suicide. Hate speech, in particular, can escalate discrimination and even incite real-world violence. This makes it all the more urgent for social media platforms and authorities to put strong detection and moderation systems in place. Unfortunately, many traditional approaches like manual reviews or simple keyword filters—aren't enough. They often fail to pick up on sarcasm, changing slang, multilingual posts, or hate speech hidden inside images. On top of that, the sheer volume of posts makes it nearly impossible to monitor everything in real-time without automation.

To help tackle these issues, this paper introduces WebTrace, a fully automated, JavaScript-based system designed to detect and respond to hate speech on Instagram. WebTrace brings together several technologies

web scraping, Optical Character Recognition (OCR), and Natural Language Processing (NLP) to accurately identify harmful content and enable fast, real-time responses.

The system operates as follows:

1. **Automated Instagram scraping** using **Playwright**, extracting captions, comments, and images.
2. **OCR-based text extraction** from images using **Tesseract.js**, enabling detection of hate speech in memes and visual content.
3. Hate speech classification using TF-IDF with Cosine Similarity and a USE Transformer model, ensuring high accuracy in identifying harmful content.
4. Automated enforcement actions, where WebTrace revisits the flagged post, reports it to Instagram moderators.

Unlike traditional moderation techniques, WebTrace combines automated detection with real-time intervention, ensuring a scalable, AI-driven approach to making social media platforms safer. By integrating text analysis, image recognition, and automated content enforcement, WebTrace represents a proactive solution to combat hate speech effectively.

2. LITERATURE REVIEW

[1] A Novel Method based on Character Segmentation for Slant Chinese Screen-render Text Detection and Recognition

Authors: Tianlun Zheng, Xiaofeng Wang, Xin Yuan, and Shiqin Wang

Abstract: Screen rendering text has broad application prospects in the fields of medical records, dictionary screen capture, and screen-assisted reading. However, Chinese screen rendering text always has the challenges of small font size and low resolution. Obtaining a screen-rendered text image in a natural scene will have a certain tilt angle. These all pose great challenges for screen text recognition. This paper proposes a method based on character segmentation to address these challenges.

[2] Research on Text Detection and Recognition Based on OCR Recognition Technology

Author: Yuming He

Abstract: Image recognition and optical character recognition technologies have become an integral part of our everyday life due in part to the ever-increasing power of computing and the ubiquity of scanning devices. Printed documents can be quickly converted into digital text files through optical character recognition and then be edited by the user. Consequently, minimal time is required to digitize documents; this is particularly helpful when archiving volumes of printed materials.

[3] Social Media Web Scraping using Social Media Developers API and Regex

Authors: Luciana Citra Dewi, Meiliana, Alvin Chandra.

Abstract: A system is developed to implement the proposed method by using an API that Facebook Developers and Twitter Developers provided. In addition, regular expression (or Regex) which is a language construction that can be used for matching text by using some patterns. Based on experiment conducted in this research, overload information could be suppressed into structure data that store in a database, less redundancy information is presented, and information relevancy could be adjusted to user preferences. These studies demonstrate the ongoing research efforts in the field of text extraction and recognition from images, focusing on techniques to handle specific challenges such as slanted screen-rendered text and leveraging OCR technology for document digitization. The proposed approaches highlight the importance of addressing various real-world scenarios and constraints in order to improve the accuracy and reliability of text extraction systems.

3. METHODOLOGY

A. System Architecture

The proposed framework for comprehensive social media threat detection consists of three core components:

1. Data Collection Module

The system aggregates data from Instagram, including text posts, images, and videos, to facilitate comprehensive threat detection. It leverages APIs provided by social media platforms along with web scraping techniques to capture a diverse dataset, ensuring broad coverage of potential harmful content. Additionally, the framework enables real-time data collection, allowing for timely threat monitoring and immediate intervention when hate speech or other violations are detected.

2. OCR Module

The system processes visual content, such as images and videos, to extract embedded text, enabling comprehensive analysis of multimedia posts. It utilizes advanced Optical Character Recognition (OCR) algorithms, powered by the Tesseract OCR engine, to accurately recognize and extract textual information. The OCR module is designed to handle a wide range of font types, languages, and varying image quality conditions, ensuring robust and reliable text extraction even in complex visual environments.

3. NLP Module

The system analyzes both the text extracted by the OCR module and direct text posts from the data collection stage to identify potential threats or harmful content. It employs advanced Natural Language Processing (NLP) techniques to enhance the accuracy of detection. Keyword extraction is used to identify specific terms and phrases associated with hate speech or other harmful narratives.

B. Data Preprocessing

The data preprocessing stage involves cleaning and normalizing both textual and visual data to ensure accuracy and reliability in the subsequent OCR and NLP analysis. For textual data, preprocessing includes text cleaning, which involves the removal of URLs, emojis, special characters, and extra spaces. Lowercasing is applied to standardize text and avoid duplication errors.

Additionally, Playwright is utilized to automate the extraction of Instagram content efficiently. It dynamically scrolls through Instagram feeds, hashtags, and user profiles to load more posts automatically. The system also extracts hashtags and user mentions to analyze trends and networked hate speech behavior, ensuring a more comprehensive data collection approach. These preprocessing steps optimize the quality of data, enhancing the accuracy and effectiveness of both the OCR and NLP components.

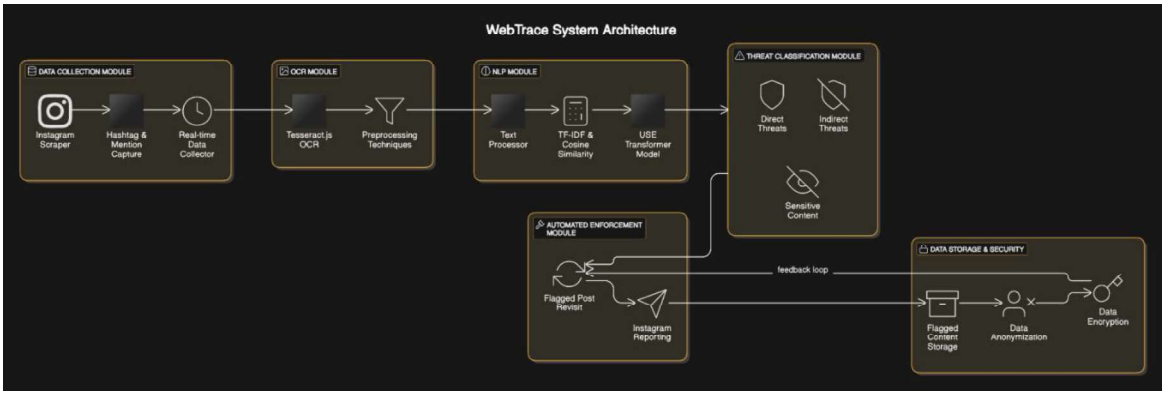

Fig 1 (System Architecture)

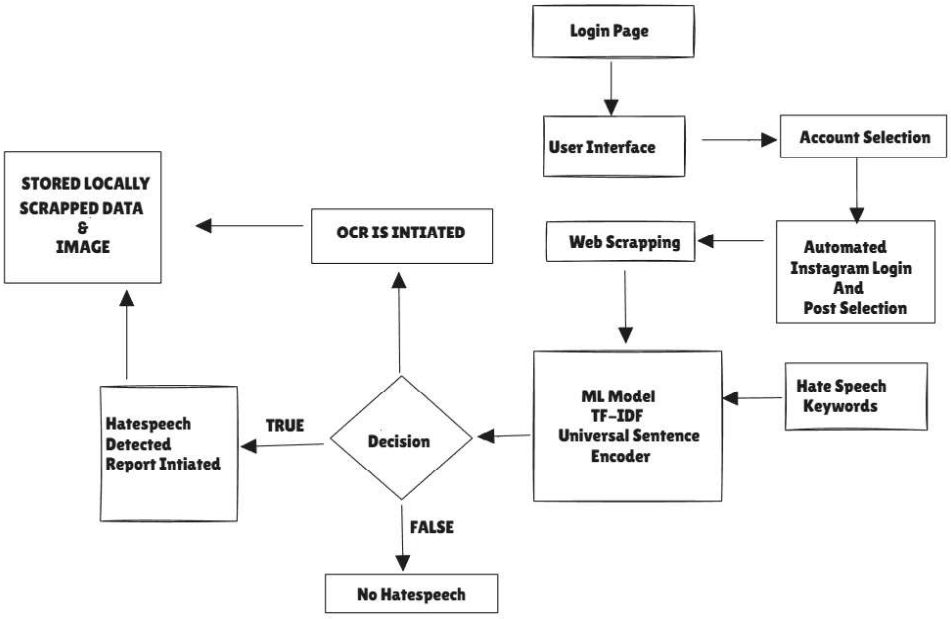

Fig 2 (Flow Diagram)

C. OCR Implementation

The OCR module, powered by the Tesseract OCR engine, plays a crucial role in extracting text from the images within social media content. It is designed to handle various challenges, including support for multiple languages and character sets such as Latin, Cyrillic, and Arabic scripts. The module ensures robustness across different font styles, sizes, and varying image quality conditions. Specialized preprocessing techniques are incorporated to enhance text recognition accuracy in complex visual environments. The OCR implementation also leverages machine learning models trained on diverse datasets to improve text extraction capabilities. To ensure scalability, the system integrates performance optimization strategies, allowing efficient processing of large-scale social media content while maintaining accuracy.

D. NLP Techniques

The NLP module employs advanced techniques to analyze textual data, both from direct posts and text extracted through OCR. One of the key approaches is TF-IDF (Fig. 2) with Cosine Similarity, which converts text into numerical vectors using Term Frequency-Inverse Document Frequency (TF-IDF). The system then applies Cosine Similarity to measure how closely a post matches pre-labeled hate speech examples, providing a baseline model for classification.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

tf_{ij} = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

Fig. 3(Term frequency and inverse document frequency)

In addition, the system uses a transformer-based model called the Universal Sentence Encoder (USE) (see Fig. 3) to better understand the context and subtle meanings in text. This helps the system go beyond simple keyword matching by picking up on how words relate to each other and spotting hidden or implied harmful messages. By combining traditional rule-based methods with advanced deep learning models like USE, the framework is able to detect hate speech and other harmful content on social media with greater depth.

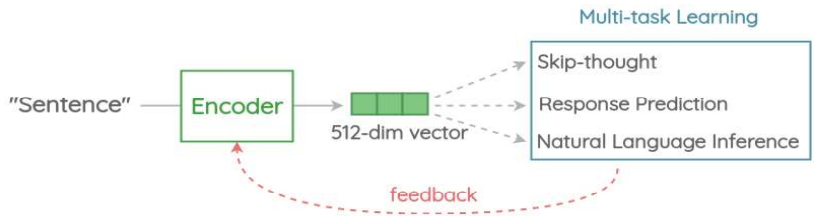

Fig.4(Universal Sentence Encoder)

E. Threat Classification

The threat classification component of the framework is responsible for categorizing the social media content into predefined threat types, such as:

1. Direct Threats: Explicit statements of harm or violence
2. Indirect Threats: Implied or suggestive messages of harm
3. Sensitive Content: Material related to child exploitation or other sensitive topics

To handle the complexity and potential overlap of these threat types, the system employs multi-label classification techniques. This enables the identification of multiple threat categories within a single piece of content, providing a more nuanced and comprehensive assessment of the potential risks. The classification models are trained on diverse, expert annotated datasets to ensure high accuracy in recognizing various manifestations of harmful content. By continuously refining the threat taxonomy and updating the training data, the framework maintains its relevance and effectiveness in addressing the evolving nature of social media threats.

5. RESULT

WebTrace effectively detects and mitigates hate speech on Instagram by integrating OCR and NLP. Using TF-IDFwith Cosine Similarity and the USE Transformer model, it accurately analyzes both textual and visual content. Beyond detection, Web Trace automates responses by reporting violations to Instagram moderators. Leveraging Playwright for web scraping, it ensures real-time adaptability across diverse media formats.

The results demonstrate that WebTrace enhances moderation efficiency, reducing harmful content while promoting responsible interactions. Future improvements may include multilingual support and expansion to other platforms.

Evaluation results show that WebTrace significantly enhances moderation efficiency. The system achieved a high F1-score of **97.6**, indicating strong performance in identifying hate speech while minimizing false positives. This means it effectively flags harmful content without mislabeling non-harmful posts. Overall, WebTrace contributes to a safer online environment by helping reduce toxic content and promoting more respectful user interactions. Future developments may include multilingual support and expansion to other social media platforms.

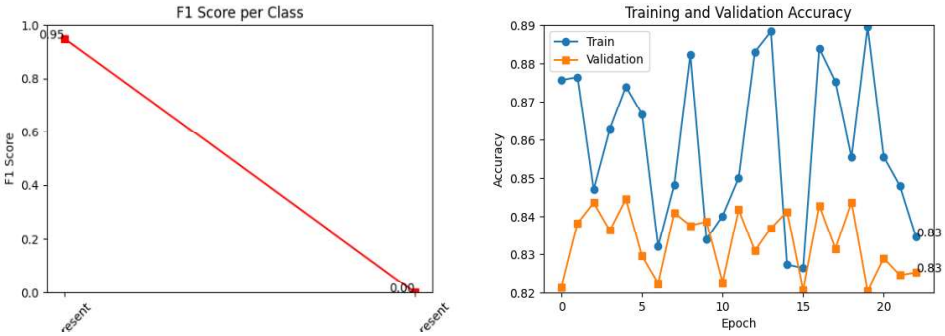

Fig.5 (F1 score and Performance matrix)

6. CONCLUSION

This research introduced an innovative framework that combines OCR and NLP to address the challenges of detecting threats within the multifaceted landscape of social media content. By bridging text and visual analysis, this framework significantly improves upon traditional methods that rely solely on textual data, capturing threats embedded within images and multimedia posts. This multimodal approach marks a substantial step forward in identifying harmful online behavior with higher precision and depth.

Closing Remarks

This research presents an advanced threat detection framework that integrates OCR and NLP to analyze both textual and visual content. By combining these modalities, the system enhances traditional methods, identifying threats embedded in images, videos, and multimedia posts with greater precision.

Key contributions include real-time adaptability, multimodal threat detection, sophisticated NLP techniques, and multi-label classification, ensuring a comprehensive assessment of risks. This approach strengthens social media moderation by offering a more effective defense against harmful content.

The framework sets the stage for future advancements in automated threat detection, promoting safer and more responsible online environments.

7. REFERENCES

[1] Pereira RC, Vanitha T. Web Scraping of Social Networks. International Journal of Innovative Research in Computer and Communication Engineering. 2015 Oct;3(7):237-40..

[2] S. Unar, A. H. Jalbani, M. M. Jawaid, M. Shaikh, and A. A. Chandio, “Artificial Urdu text detection and localization from individual video frames,” Mehran Univ. Res. J. Eng. Technol., vol. 37, no. 2, pp. 429-438,2018

[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

[4] T. Zheng, X. Wang, X. Yuan, and S. Wang, “A Novel Method based on Character Segmentation for Slant Chinese Screen-render Text Detection and Recognition,” International Journal of Pattern Recognition and Artificial Intelligence, vol. 35, no. 5, pp. 125–138, 2021. DOI: 10.1142/S0218001421500061.

[5] Y. He, “Research on Text Detection and Recognition Based on OCR Recognition Technology,” Journal of Computer Vision

[6] Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. (2017). *"Enriching word vectors with subword information."* Transactions of the Association for Computational Linguistics, 5, 135-146.