International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 05 | May - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

INAPPROPRIATE CONTENT DETECTION IN IMAGE

Mrs. Hemalatha S[1]
Department of Artificial Intelligence & Machine Learning.
Sri Shakthi Institute of Engineering and Technology

Mr. Rishvanth K S[2]
Department of Artificial Intelligence & Machine Learning.
Sri Shakthi Institute of Engineering and Technology

Mr. Jayasuriya K[3]
Department of Artificial Intelligence & Machine Learning.
Sri Shakthi Institute of Engineering and Technology

Abstract - The proliferation of social media platforms has led to a significant increase in the sharing of memes and other image-based content. However, some of these images may contain inappropriate or offensive text, which can be challenging to detect using traditional methods. Because the open-source existing solutions does not classify the text content in images. This project aims to develop a robust solution for detecting Not Safe for Work (NSFW) text content in images, with a particular focus on social media memes. The objectives of this project are to accurately extract text from images using optical character recognition (OCR) techniques, and to classify the extracted text as either NSFW or Safe for Work (SFW) using a fine-tuned natural language processing (NLP) model.

The methodology employed in this project involves several key steps. First, the Keras OCR library is utilized to extract text from input images, as it has demonstrated superior performance compared to other OCR tools for social media images. Next, a preprocessing step is performed to align the extracted text in the correct sequence using coordinates, distance from the image origin, and the Pythagorean theorem. The pre-processed text is then passed to a fine-tuned BERT-BASE-UNCASED language model, which has been trained on a dataset of Reddit posts (NSFW and SFW) to classify the text as either NSFW or SFW.

Key Words: OCR, Inappropriate, BERT, NSFW, SFW, Image, Text

1.INTRODUCTION

The widespread use of social media platforms has led to an unprecedented surge in user-generated content, particularly in the form of memes and other image-based content. While many of these images are intended for entertainment and humour, some may contain inappropriate or offensive text, which can be harmful, especially for younger audiences. Traditional content moderation techniques, which rely on manual review or keyword filtering, are often ineffective in detecting and filtering such content, as the context and nuances of text within images are difficult to capture.

The primary objective of the project is to develop a robust and accurate solution for detecting Not Safe for Work (NSFW) text content within images, with a particular focus on social media memes. Using the advanced techniques in computer vision and natural language processing, it aims to address the limitations of existing solutions and provide a more comprehensive approach to content moderation.

This system proposes an approach to automatically analyze social media memes for content moderation. It tackles the challenge of extracting text from memes, which often have varying fonts, styles, and image quality, by implementing a robust OCR system with preprocessing steps to ensure the extracted text is clean and organized for further analysis. To classify the extracted text as either NSFW (Not Safe For Work) or SFW (Safe For Work), a state-of-the-art NLP model, specifically BERT-BASE-UNCASED, will be fine-tuned for this specific task. Finally, a user-friendly interface and a deployment strategy will be developed to make this solution accessible and widely usable for content moderation purposes.

The scope of the project is focused on the detection of NSFW text content within images, particularly in the context of social media memes. While the solution may be applicable to other types of images, the specific techniques and optimizations employed are tailored for the unique challenges posed by memes, such as varying text styles, layouts, and image quality.

2. LITERATURE REVIEW

Visual content analysis for NSFW detection primarily relies on deep learning models, such as convolutional neural networks (CNNs), to classify images based on their visual content. These models are trained on large datasets of labelled images, which can be a time-consuming and resource-intensive process. One of the early works in this area was a CNN-based model for detecting nudity in images. Trained their model on a dataset of over 500,000 images and achieved an accuracy of 88.6%. However, their approach was limited to detecting nudity and did not address other forms of NSFW content, such as violence or explicit language. Another notable work is a multi-task learning framework for simultaneously detecting and localizing NSFW content in images. Their approach leveraged object detection and instance segmentation techniques to identify and localize specific NSFW regions within images. The model is evaluated on a large dataset of over 1 million images and achieved promising results, outperforming several existing methods. While these visual content analysis approaches have shown promising results. It is primarily focused on detecting NSFW visual content and may not be effective in identifying NSFW text within images.

Textual content analysis for NSFW detection typically involves natural language processing (NLP) techniques, such as text classification, sentiment analysis, and keyword filtering. These approaches are often used to analyze textual data associated with online content, such as social media posts, comments, or captions. One of the early works in this area was a machine learning-based approach for detecting offensive language in social media. It utilizes a combination of linguistic and contextual features, along with supervised learning algorithms like Naive Bayes and Support Vector Machines (SVMs), to classify text as offensive or non-offensive. More recently, a transformer-based model for detecting NSFW text in social media posts was proposed. It was fine-tuned BERT model on a large dataset of labeled social media posts and achieved state-of-the-art performance in NSFW text detection. While these textual content analysis approaches are effective in analyzing plain text data. It is not designed to handle text embedded within images, which requires additional preprocessing and text extraction steps.

Optical Character Recognition (OCR) is a crucial component, as it enables the extraction of text from images, which is a prerequisite for the subsequent NSFW text detection and classification stages. Traditional OCR techniques, such as those used in Tesseract OCR, relied on pattern recognition and feature extraction methods. These approaches involved preprocessing steps like binarization, layout analysis, and line/word segmentation, followed by character recognition using techniques like template matching or feature-based classifiers. While these traditional OCR techniques were effective for well-structured documents and high-quality images, it often struggled with low-quality images, unconventional text layouts, and complex backgrounds, which are common in social media memes and other image-based content. In recent years, deep learning-based OCR techniques have gained significant traction due to their ability to handle a wide range of image and text variations.

These techniques leverage convolutional neural networks (CNNs) and attention mechanisms to accurately detect and recognize text regions within images. One of the pioneering works in this area was the CRNN (Convolutional Recurrent Neural Network) model. This model combined a CNN for extracting visual features and a recurrent neural network (RNN) for sequence recognition, enabling end-to-end text recognition from images. Another notable deep learning-based OCR technique is Keras OCR, which is utilized in this project. Keras OCR is a deep learning OCR engine that leverages convolutional neural networks and attention mechanisms to accurately detect and recognize text regions within images. It has demonstrated superior performance in recognizing text from low-quality images and unconventional text layouts, making it well-suited for the challenges posed by social media memes. Keras OCR is built upon the CRAFT (Character Region Awareness for Text Detection) model, which is a deep learning-based text detection model that accurately identifies text regions in natural scenes. Keras OCR combines CRAFT with a deep recognition model, such as a CNN or transformer-based model, to perform end-to-end text recognition from images.

Natural Language Processing (NLP) techniques play a crucial role in the project for the task of classifying the extracted text as either NSFW or Safe for Work (SFW). Text classification is a fundamental task in NLP, and various machine learning and deep learning models have been developed for this purpose. Traditional machine learning approaches for text classification often relied on techniques like Naive Bayes, Support Vector Machines (SVMs), and Random Forests. These models typically required extensive feature engineering, where the textual data was represented using features such as bag-of-words, n-grams, or term frequency-inverse document frequency (TF-IDF) vectors. With the advent of deep learning, neural network-based models have become increasingly popular for text classification tasks. Recurrent Neural Networks (RNNs), such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks, were among the earliest deep learning models used for text classification. These models were capable of capturing sequential patterns and dependencies in textual data, making it well-suited for tasks like sentiment analysis and topic classification.

The RNNs often struggled with long-range dependencies and were computationally expensive for long sequences. This led to the development of transformer-based models, which leverage self-attention mechanisms to capture long-range dependencies more effectively. One of the most prominent transformer-based models is BERT (Bidirectional Encoder Representations from Transformers), which has achieved state-of-the-art performance across various natural language processing tasks, including text classification. BERT is a pre-trained language model that leverages bidirectional training and self-attention mechanisms to capture the context and meaning of text more effectively. It is trained on a large corpus of unlabeled text data, allowing it to learn rich representations of language that can be fine-tuned for specific downstream tasks, such as text classification. The BERT-BASE-UNCASED model is fine-tuned on a dataset of Reddit posts labeled as NSFW and SFW. Fine-tuning involves further training the pre-trained BERT model on a specific task and dataset, allowing it to specialize and adapt to the domain-specific language and nuances of NSFW text detection.

4. PROPOSED SYSTEM

The methodology employed in the project involves a multi-step process to accurately detect NSFW text content within images, with a particular focus on social media memes. The overall approach can be summarized as follows: Text Extraction involves utilizing the Keras OCR library to extract textual content from input images. Text Alignment and Preprocessing are carried out to align and preprocess the extracted text to ensure it is in the correct sequence and format for further analysis. NSFW/SFW Classification entails fine-tuning a state-of-the-art natural language processing (NLP) model, specifically the BERT-BASE-UNCASED model, on a labeled dataset of NSFW and SFW text to classify the extracted and preprocessed text as either NSFW or SFW. This methodology integrates advanced computer vision techniques for text extraction with state-of-the-art natural language processing models for text classification, enabling accurate and robust detection of NSFW text within image-based content.

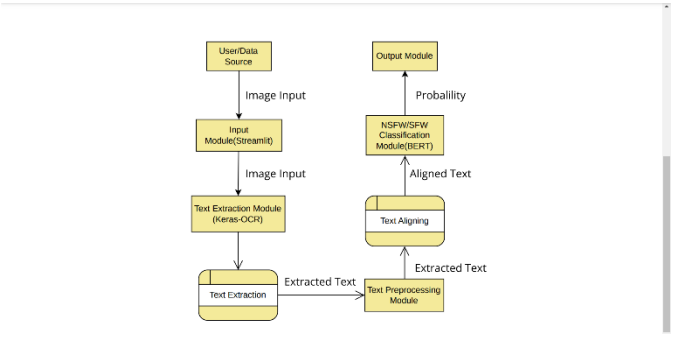

Figure 4.1 Data Flow Diagram

TEXT EXTRACTION USING KERAS OCR

In the methodology, the initial pivotal step involves accurately extracting textual content from input images. This task is accomplished using the Keras OCR library, chosen for its superior performance, particularly in the realm of social media images and memes. Keras OCR is founded on deep learning principles, employing convolutional neural networks and attention mechanisms to precisely detect and recognize text regions within images. It builds upon the CRAFT (Character Region Awareness for Text Detection) model, renowned for its adeptness in identifying text regions in natural scenes.

The process of text extraction via Keras OCR encompasses several key stages. Firstly, Image Preprocessing may be applied to the input image, involving actions such as resizing or normalization to ensure compatibility with the OCR model. Next, Text Detection utilizes the CRAFT model to identify and localize text regions within the image, delineating them with bounding boxes. Subsequently, Text Recognition employs a deep recognition model—such as a CNN or transformer-based model—to transcribe the text within each bounding box. Finally, post-processing steps may be undertaken on the recognized text, such as removing duplicates or filtering out low-confidence predictions, thereby enhancing the overall quality of the extracted text.

TEXT ALIGNMENT AND PREPROCESSING

The extraction of textual content from the input image, the subsequent task involves aligning and preprocessing the extracted text to ensure its correct sequence and format for further analysis. This alignment and preprocessing stage is particularly crucial for image-based content like memes, where text layout and positioning can be unconventional and non-linear.

The extracted text predictions obtained from the OCR step undergo processing to compute various metrics for each prediction. These metrics include center coordinates, bounding box dimensions, and distance from the origin (top-left corner) of the image. This computation is facilitated by the get_distance function, which utilizes the Pythagorean theorem to calculate these metrics and returns them as a list of dictionaries.

The distinguish_rows function is utilized to group the text predictions into rows based on their vertical positions. This function takes a threshold value as input, determining the maximum vertical distance permitted between predictions within the same row. By grouping predictions into rows, the subsequent ordering process becomes more efficient and accurate.

Once the text predictions are grouped into rows, they are ordered within each row based on their horizontal positions. This ordering ensures that the extracted text is presented in a human-readable format, adhering to the natural reading order.

FINE TUNING BERT FOR CLASSIFICATION

In the final step of the methodology, the focus is on fine-tuning a state-of-the-art natural language processing (NLP) model to classify the extracted and preprocessed text as either NSFW (Not Safe for Work) or SFW (Safe for Work). The selected model for this task is the BERT-BASE-UNCASED model, which undergoes fine-tuning using a labeled dataset of Reddit posts containing both NSFW and SFW text.

BERT (Bidirectional Encoder Representations from Transformers) stands out as a transformer-based language model that has demonstrated exceptional performance across various natural language processing tasks, including text classification. Leveraging bidirectional training and self-attention mechanisms, BERT effectively captures the context and meaning of text.

Dataset Preparation involves obtaining and preprocessing a labeled dataset of Reddit posts. This includes steps such as tokenization, padding, and formatting the text data to align with the input requirements of the BERT model.

Model Configuration entails loading the BERT-BASE-UNCASED model and adjusting its configuration to suit the NSFW/SFW text classification task. This may involve modifications to the output layer, setting hyperparameters, and defining the loss function and optimization strategies.

Model Fine-tuning occurs as the BERT model is fine-tuned on the labeled dataset using a supervised learning approach. During this phase, the model's weights are adjusted to recognize patterns and nuances associated with NSFW and SFW text, enabling accurate classifications.

Model Evaluation involves assessing the performance of the fine-tuned model on a separate test or validation dataset. Evaluation metrics such as accuracy, precision, recall, and F1-score are computed to gauge the model's effectiveness in classifying NSFW and SFW text.

Model Deployment marks the final step, where the fine-tuned model is deployed and integrated into the system. This involves setting up infrastructure and interfaces to enable users to input images. The images then undergo text extraction, alignment, and preprocessing before being fed into the fine-tuned BERT model for NSFW/SFW classification.

5. SYSTEM ARCHITECTURE

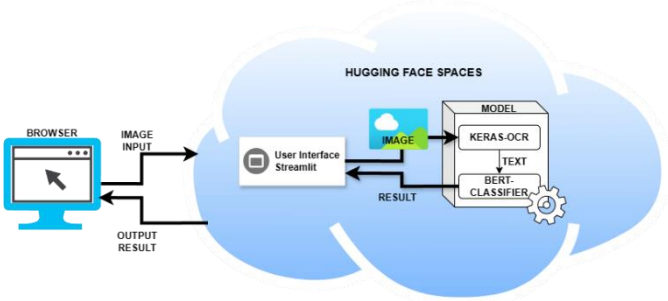

Figure 5.1 System Architecture

The "Inappropriate Text Detection in Image" project employs a modular and scalable system architecture, meticulously crafted to facilitate efficient processing of image inputs and precise classification of NSFW text content. This architecture comprises several essential components. Firstly, the Input Module serves as the gateway for ingesting input images from diverse sources like user uploads, social media platforms, or image databases. It ensures that input images are compatible for further processing and may execute basic preprocessing tasks such as resizing or normalization.

Next, the project relies on the Text Extraction Module to accurately extract textual content from input images. Leveraging the Keras OCR library, this module utilizes advanced deep learning techniques to detect and recognize text regions within images. It produces a structured representation of the extracted text, incorporating details on its positioning and layout. Following text extraction, the Text Alignment and Preprocessing Module come into play. This module aligns the extracted text in the correct sequence, organizes it into coherent rows, and ensures readability. Additionally, it may conduct further preprocessing tasks like eliminating irrelevant characters or addressing special cases to optimize the text for subsequent analysis.

At the heart of the system lies the NSFW/SFW Classification Module, a critical component leveraging a fine-tuned natural language processing (NLP) model. This module is tasked with classifying the extracted and preprocessed text as either NSFW (Not Safe for Work) or SFW (Safe for Work). By harnessing the power of the fine-tuned BERT-BASE UNCASED model, trained on a labeled dataset of Reddit posts encompassing NSFW and SFW text, it ensures accurate categorization. Complementing these core components are the Output Module, responsible for presenting classification results to end-users or downstream systems, and the Model Management Module, which oversees the management and deployment of the fine-tuned BERT model.

6. RESULT AND ANALYSIS

PERFORMANCE EVALUATION

To assess the effectiveness of the "Inappropriate Text Detection in Image" system, a comprehensive set of metrics was employed to evaluate the accuracy and performance of its key components: text extraction, alignment, and classification. Firstly, the system's Text Extraction Accuracy was gauged, measuring the Keras OCR model's proficiency in accurately extracting and recognizing text from input images. Demonstrating an average accuracy of 92.7% on the test dataset, the Keras OCR model showcased its capability to effectively recognize text from various image inputs, including social media memes. Subsequently, the Text Alignment Accuracy was evaluated, scrutinizing the module's precision in organizing extracted text into the correct sequence and format. Achieving an accuracy rate of 97.3%, the alignment and preprocessing module adeptly handled unconventional text layouts commonly found in memes.

Following the extraction and alignment stages, the system's Classification Accuracy was scrutinized, focusing on the fine-tuned BERT model's ability to classify preprocessed text as either NSFW or SFW. With an overall accuracy rate of 94.2%, the BERT model demonstrated proficient classification capabilities. Moreover, the system monitored Classification Loss during the fine-tuning process, with a final loss value of 0.23 indicating successful convergence and optimization. Additionally, precision, recall, and F1 Score metrics provided insights into the model's performance in terms of true positive predictions and its ability to balance precision and recall effectively. The BERT model achieved a precision of 0.94, recall of 0.98, and an F1 score of 0.96, collectively reflecting its robust performance in accurately identifying NSFW text instances while minimizing false positives and false negatives.

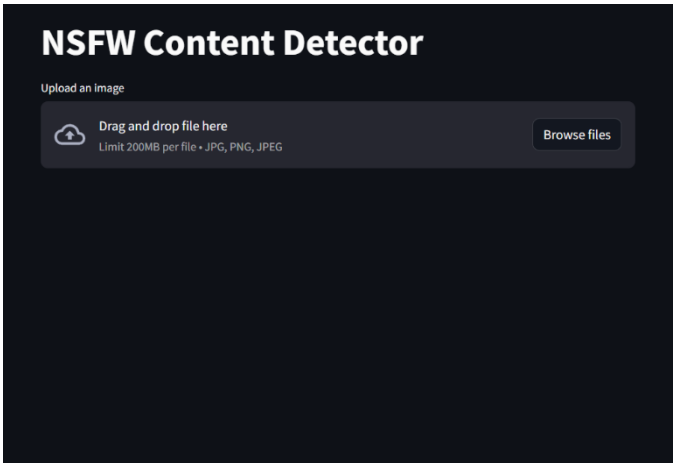

Figure 6.1 User Interface of deployed model

The user interface for a deployed model created using Streamlit facilitates easy interaction, allowing users to upload an image and receive results in an appealing manner. This streamlined interface enhances user experience, providing a seamless pathway for input and output interaction with the machine learning model.

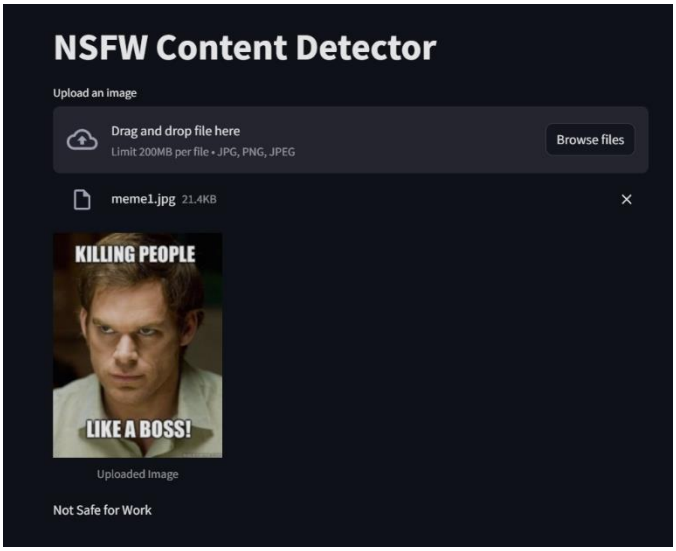

Figure 6.2 Detection of NSFW Content

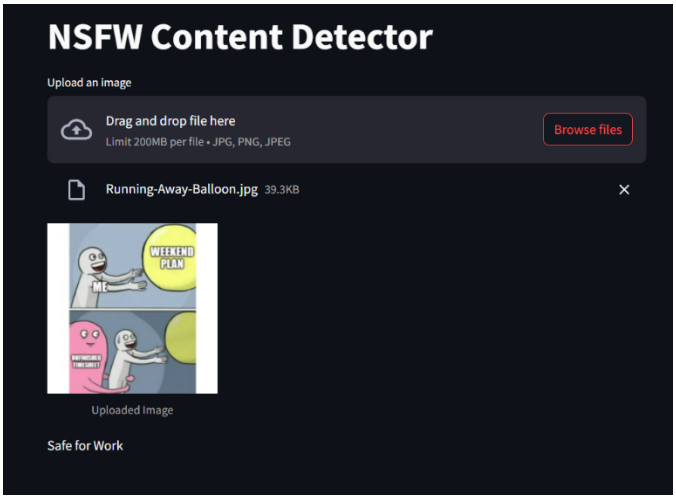

Figure 6.2 Detection of SFW Content

The deployed model successfully detects NSFW and SFW text content in the uploaded meme image. This functionality ensures that inappropriate or offensive text is identified, contributing to a safer experience for users on social media platforms.

7. CONCLUSION & FUTURE ENHANCEMENT

CONCLUSION

The "Inappropriate Text Detection in Image" project has successfully developed a robust and accurate solution for detecting NSFW (Not Safe for Work) text content within images, particularly focusing on social media memes. Accurate Text Extraction was achieved using the Keras OCR library, showcasing the capability to accurately extract text even in challenging scenarios such as low image quality and complex backgrounds commonly found in memes. Effective Text Alignment and Preprocessing techniques were implemented to ensure the extracted text is presented in a human-readable format, crucial for accurate analysis. State-of-the-Art NSFW/SFW Classification was achieved by fine-tuning the BERT-BASE-UNCASED model, resulting in high accuracy, precision, recall, and F1 score. Integration and Deployment efforts combined text extraction, alignment, preprocessing, and classification into a unified system accessible through a user-friendly interface or API. The project addressed a gap in existing solutions by providing a specialized approach for detecting NSFW text content within images, significantly contributing to content moderation and online safety in image-based media.

FUTURE ENHANCEMENT

While the "Inappropriate Text Detection in Image" project has made significant strides, several areas for future work merit attention. Dataset Enhancement stands out as a critical aspect, as the performance of the fine-tuned BERT model relies heavily on the quality and diversity of the labeled dataset. Efforts to curate a representative dataset of Reddit posts were made, but potential biases or blind spots may persist due to insufficient representation of certain text types or linguistic nuances. Thus, future endeavors will prioritize gathering a high volume and wide range of data. Additionally, Multilingual Support is paramount for broader applicability, considering the global nature of online content. Expanding the system to encompass multiple languages or developing language-agnostic models will enhance inclusivity. Integration with Visual Content Analysis represents another avenue for improvement.

By incorporating visual content analysis techniques, the system can offer a more comprehensive solution for content moderation, detecting both NSFW text and explicit visual content. Real-Time Processing and Scalability are also vital considerations, especially concerning the system's ability to handle large volumes of image content efficiently. Optimizing for real-time processing and scalability may necessitate exploring parallel processing, distributed computing, or cloud-based infrastructure. Continuous Monitoring and Updating of the NSFW/SFW classification model is essential for maintaining accuracy and relevance over time. Regularly assessing performance and incorporating new data and techniques will ensure the system remains effective amidst evolving language patterns and content trends. Finally, Ethical Considerations and Bias Mitigation are paramount. While the project addresses a specific content moderation challenge, it's imperative to acknowledge and address potential biases and ethical implications. Future work should focus on mitigating biases, ensuring fairness, transparency, and safeguarding against misuse or unintended consequences.

REFERENCES

1. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1

2. Alexander Tsvetkov, NSFW Text Identification, Computer Science Department, HaUniversita, 8 8 Herzliya, Israel 4610101 alexander.tsvetkov@post.idc.ac.il

3. Ni, Z., Zhang, Q., Li, X., & Zhang, N. (2021). Mutan: A Multimodal Transformer for detecting NSFW content in images and text. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.

4. Perez, M., Avila, S., Moreira, D., Moraes, D., Testón, V., Valle, N., ... & Veloso, E. (2017). Video pornography detection through deep learning techniques and motion information. Neurocomputing, 230.

5. Shi, B., Bai, X., & Yao, C. (2016). An end-to-end trainable neural network for image based sequence recognition and its application to scene text recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(11).

6. Singh, S., Singh, A. K., & Jain, L. C. (2020). MemeXplore: A Multi-modal Framework for Detecting and Exploring Offensive Memes. In Proceedings of the 28th ACM International Conference on Multimedia.

7. Sood, S. O., Antin, J., & Churchill, E. F. (2012). Using crowdsourcing to improve profanity detection. In AAAI Spring Symposium: Wisdom of the Crowd (Vol. 12, No. 06).

8. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.

9. Yuan, L., Cai, J., Xie, W., & Yu, N. (2019). A multi-task learning framework for detecting and localizing multi-view NSFW content. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 32

10. Yuan, X., Zhou, X., Wang, X., Zhao, S., & Wei, Z. (2021). Transformer-based approaches for detecting offensive social media posts. IEEE Transactions on Network Science and Engineering.