Meme Based Cyberbullying Detection

IVP Project Report

Rajat Srivastava[1,2], Nikunj Gupta[1,3], Kaushik Mullick[1,4], Viraj Jagtap[1,5], Kuber Jain[1,6], Dhairya Bhadani[1,7] and Sanyam Agrawal[1,8]

[1] Indian Institute of Information Technology, Prayagraj, Uttar Pradesh 211012, India
[2] iit2021109@iiita.ac.in, [3] iit2021166@iiita.ac.in, [4] iit2021168@iiita.ac.in,
[5] iit2021170@iiita.ac.in, [6] iit2021184@iiita.ac.in, [7] iit2021192@iiita.ac.in,
[8] iit2021207@iiita.ac.in

Abstract. The rise of cyberbullying and hate speech embedded within memes in the digital landscape necessitates innovative solutions. In response, our research presents a multimodal algorithm for cyberbullying detection in memes. By combining BERT's linguistic understanding with VGG-16's image analysis, the algorithm extracts text through Optical Character Recognition and relevant image features. These modalities are merged into a high-dimensional feature vector, enabling a deeper comprehension of memes' content, with an emphasis on context. While the algorithm shows promise, it faces challenges in understanding nuanced language, resource-intensive processing, and potential biases. However, it underscores the importance of diverse training data, regular updates, user feedback, and ethical considerations in fostering a safer and more inclusive online community.

Keywords: Cyberbullying, Hate speech, Meme analysis, Multimodal approach, Text-image fusion, BERT model, VGG-16 CNN, Optical Character Recognition (OCR), Image feature extraction, Online community, Context analysis, Ethical considerations, Memes and context, Online safety, Online communication, Linguistic understanding, Visual analysis, Algorithm for hate speech detection, User feedback, Training data diversity.

1 Introduction

1.1 What are memes:-

"Memes" are typically humorous or satirical cultural artifacts that traverse the digital realm with notable velocity. They manifest in diverse formats, including images, videos, and textual constructs, all subject to adaptation and contextual reconfiguration. Internet memes encompass a wide spectrum of subjects, from commentary on contemporary socio political issues to references rooted in popular culture. They furnish a medium through which denizens of the internet engage with and comment upon various facets of contemporary society, encapsulating multifaceted expressions and insights within a compact digital framework.[11]

1.2 Cyberbullying through memes:-

Cyberbullying is a form of harassment and aggression conducted through digital means, encompassing various online platforms and communication channels. Cyberbullying, illustrated through memes, is the deployment of digital humor in a manner that maliciously targets individuals with the intention of causing emotional distress, social harm, or reputation damage. Memes, often employed for humorous or satirical purposes, are repurposed to perpetrate hurtful, demeaning, or harmful narratives against specific individuals or groups in the online sphere. These manipulative visuals and messages can encompass personal attacks, derogatory comments, false allegations, or the unauthorized sharing of sensitive information, all disseminated widely on digital platforms.[10]

Fig. 1. The picture on the left demonstrates a meme which would not be considered as cyberbullying.[17] The picture on the right demonstrates a meme which would be considered as cyberbullying.[16]

2 Literature Review - Hate Speech in Pixels:Detection of Offensive Memes towards Automatic Moderation

This paper proposes a multimodal approach for detecting hate speech in internet memes by combining visual and linguistic representations. The authors built a dataset of 5,020 memes and found that the visual modality is more informative than the linguistic one for hate speech detection. However, they acknowledge that the task is still challenging and human moderation is necessary. The authors provide code and models for reproducibility. The paper references other works on hate speech and pornography detection in text and social media.

2.1 Hate Speech Detection in Social Networks:-

Previous research on hate speech detection in text-based social media platforms has focused on using machine learning techniques to analyze linguistic features like keywords, n-grams, and sentiment analysis. However, text-based detection methods face challenges due to the dynamic nature of hate speech, the use of sarcasm and coded language, and the lack of context. Hate speech can also be disguised or subtly expressed through euphemisms and metaphors. While text-based detection methods have made progress, they still struggle to capture the nuances of hate speech.

2.2 Multimodal Approaches in Social Media Analysis:-

There is an emerging trend in social media analysis that involves incorporating multiple modalities, such as text and visuals, to gain a more comprehensive understanding of the content . This approach has been applied to hate speech detection, where the combination of linguistic and visual analysis has shown improved accuracy . The multimodal approach offers advantages in capturing the nuances and complexities of hate speech, as visual cues can provide additional contextual information . However, challenges such as the dynamic nature of hate speech and the use of coded language still exist . Overall, the multimodal approach shows promise in enhancing automated moderation systems in social media .

2.3 Previous Work on Hate Speech Detection:-

Previous work on hate speech detection has primarily focused on linguistic approaches, which involve analyzing the textual content of social media posts . These methods often use machine learning algorithms to extract linguistic features like keywords, n-grams, and sentiment analysis . For example, some studies have used bag-of-words or N-gram features along with expert knowledge of hate speech keywords to train binary classifiers. The strengths of linguistic approaches lie in their ability to capture explicit hate speech that is directly expressed through language. They can effectively identify instances where offensive or discriminatory terms are used. However, linguistic methods struggle with detecting hate speech that is disguised or subtly expressed, such as through sarcasm or coded language . Additionally, linguistic approaches may not capture the nuances and complexities of hate speech that involve visual cues or implicit biases.

2.4 The Significance of Visual Modality:-

Visual cues in memes, such as offensive imagery or symbols, play a crucial role in detecting hate speech compared to linguistic analysis . Analyzing visual content presents challenges, requiring advanced computer vision techniques and accurate interpretation of visual cues. The dynamic nature of memes and evolving trends in social media necessitate continuous adaptation of detection models. Considering visual cues and incorporating additional information, such as societal context, is essential for effectively detecting and moderating hate speech in memes. This shift towards the visual modality offers opportunities to improve accuracy and effectiveness in addressing hate speech in memes.

2.5 **Dataset for Training and Evaluation:-**

The research team created a dataset for detecting hate speech in memes, comprising 5,020 images. The hate memes (1,695 in total) were sourced from Google Images using specific search queries related to racial, religious, and ethnic topics. Non-hate memes (3,325) were obtained from the Reddit Memes Dataset, assuming they did not contain hate messages. The dataset was divided into a training set (4,266 memes) and a validation set (754 memes) with an equal distribution of classes. Due to limited data, a separate test set was not created, and reliance was placed on the validation set metrics for evaluation.

2.6 **Implementation Details of the Process:-**

Text Extraction: The authors used Optical Character Recognition (OCR) to extract the text from the memes. They specifically used the Tesseract 4.0.0 OCR tool, which is a widely used OCR Engine. Text Encoding: The text detected by the OCR was encoded using BERT, a state-of-the-art language representation model. The authors used a PyTorch implementation of BERT, specifically the "bert-base-multilingual-cased" version, which has 12 layers, 768 hidden dimensions, and 12 attention heads. Image Feature Extraction: The authors used a VGG-16 convolutional neural network (CNN) that was pretrained on the ImageNet dataset to extract features from the images. They used the activations from a hidden layer of the VGG-16 model, specifically the last hidden layer before the output, which has 4096 dimensions. Multimodal Representation: The text and image encodings were combined by concatenation, resulting in a multimodal feature vector of 4,864 dimensions. Hate Speech Detection: The multimodal representation was then fed as input into a multi-layer perceptron (MLP) with two hidden layers of 100 neurons each, with a ReLU activation function. The final output of the MLP was a single neuron used to predict the hate speech detection score.

2.7 **Experiments and Results:-**

The study aimed to detect hate speech in memes using a multimodal approach. The vision-only configuration outperformed the language-only configuration, highlighting the importance of visual cues. The best multimodal model achieved an average precision of 0.81. Analysis of false positives and false negatives revealed a visual bias in the dataset and challenges with OCR recognition and language encoding. While automation is possible, human moderation is still necessary. Additional information beyond the meme itself, such as societal context, should be considered. Further research is needed to address the complexities of hate speech detection in memes.

2.8 **Conclusion:-**

The study aimed to detect hate speech in memes using a multimodal approach combining visual and linguistic information. The results showed that the vision-only configuration outperformed the language-only configuration, emphasizing the importance of visual cues. The multimodal approach achieved the best results, but the improvement over the vision-only approach was relatively small. The analysis

revealed a visual bias in the dataset and challenges with OCR recognition and language encoding. The study concluded that while automation is possible, human moderation is still necessary, and additional information beyond the meme itself should be considered. Future research should focus on addressing these limitations and incorporating societal context for more effective hate speech detection in memes.

Link to LR TABLES OF Papers:
https://www.notion.so/LR-Table-90266e3c3c794b719b0beddb3b7329e3

LR Table

Paper Title	Year	Conference	Research Objective	Dataset	Methodology Used	Parameters used for Evaluating	Accuracy	Limitations of tested systems
Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text	2020	**Workshop on Trolling, Aggression and Cyberbullying (TRAC)**	Created a meme classifier to detect offensive content by essential modalities combination; leveraging 2016 U.S. presidential election memes, the MultiOFF dataset was formed, facilitating the development of an effective offensive content detection system.	https://drive.google.com/drive/folders/1hKLOtpVmF45IoBmJPwojgq6XraLtHmV6?usp=sharing	An early fusion technique is being used to combine the image and text modality, and its effectiveness is being investigated by comparing it with a text- and an image-only baseline.	Precision, Recall, and F-Score		The paper's dataset is relatively small and manually annotated, posing challenges due to limited diversity and potential bias, underscoring the importance of addressing these issues in research and analysis.
Exploring Hate Speech Detection in Multimodal Publications	2019	IEEE Winter Conference on Applications of Computer Vision (WACV)	This study focuses on hate speech detection in multimodal publications, combining text and images. Using the Twitter dataset MMHS150K, the researchers propose various models that jointly analyze textual and visual information, comparing their performance with unimodal detection methods.	https://www.kaggle.com/datasets/victorcallejasf/multimodal-hate-speech	In the Multimodal Treatment section, the authors introduce the TKM (Textual Kernel Model) for hate speech detection, employing text and image fusion. They train various models (FCM, SCM, TKM) considering input availability, analyze their contributions, and compare multimodal performance.	F-score, Accuracy and Area under ROC curve	68.2%	Challenges in hate speech detection include subjective judgment discrepancies in annotations, particularly pronounced in multimodal tasks. The intricate relations between visual and textual elements in multimodal publications, along with a limited dataset of examples, pose difficulties for neural network learning.
DISARM: Detecting the Victims Targeted by Harmful Memes	2022	Findings of the Association for Computational Linguistics: NAACL 2022	Introduce DISARM (Detecting victims targeted by harmful Memes), employing named entity recognition and person identification to identify meme entities. Our framework incorporates a contextualized multimodal deep neural network for classifying harmful intent.	https://github.com/LCS2-IIITD/DISARM/tree/main/DISARM_Dataset	The paper introduces DISARM (Detecting Entities Targeted by Harmful Memes), a multimodal deep learning framework that combines visual and textual information to detect entities targeted by harmful memes.	Accuracy, precision (0.74), recall (0.86), F1-score (0.78)	73.9%	The dataset exhibits a notable bias as it is excessively centered around the United States, potentially introducing biases owing to manual annotation, thereby limiting its representativeness and generalizability to diverse cultural and linguistic contexts.
An approach to detect offence in Memes using Natural Language Processing(NLP) and Deep learning	2021	International Conference on Computer Communication and Informatics (ICCCI -2021), Jan. 27 – 29, 2021, Coimbatore, INDIA	This paper presents an approach to detect offense in memes using Natural Language Processing (NLP) and deep learning.	https://github.com/roushangiri/Offence-Evaluation/blob/master/Meme_dataset.csv	First, it will extract the text from the given image, then it will classify the given text as offensive or not offensive. If the text is found to be offensive then in the third step it will further classify offensive text in three categories namely slight offensive, very offensive and hateful offensive.	Accuracy, Training Loss, Value Accuracy, Value Loss	93%	Cyberbullying often involves image-based harassment, where offensive images are used to target individuals or groups. Excluding the image analysis component limits the model's capability to identify and mitigate instances of image-based cyberbullying, which is a prevalent and concerning form of online harassment.
Racist or Sexist Meme? Classifying Memes beyond Hateful	2021	Proceedings of the Fifth Workshop on Online Abuse and Harms	This work presents a multimodal pipeline that takes both visual and textual features from memes into account to (1) identify the protected category (e.g. race, sex etc.) that has been attacked; and (2) detect the type of attack (e.g. contempt, slurs etc.).	https://www.workshoponlineabuse.com https://github.com/facebookresearch/fine_grained_hateful_memes https://github.com/harisbinzia/HatefulMemes	A simple four step pipeline: (i) We extract text from the meme. (ii) We embed the meme image and the text into visual and textual representations. (iii) We concatenate the visual and textual embeddings. (iv) We train a multi-label Logistic Regression classifier using scikit-learn to predict the protected category attacked in the meme and the type of attack.	Accuracy, Precision, Recall, f1		The dataset used for evaluation is imbalanced, with a large number of non-hateful memes compared to hateful ones. Imbalanced datasets can pose challenges in training models, and the paper does not thoroughly address the potential impact of this class imbalance on the model's performance and generalization to real-world scenarios.
Hate Speech in Pixels: Detection of Offensive Memes towards Automatic Moderation	2019	Universität Politecnica de Catalunya - UPC	The motivation behind this objective is the significant impact of memes, a popular form of multimedia content, in spreading hate through social networks. The goal is to develop a system that can automatically detect hate speech in memes and contribute to reducing the harmful societal impact of such content.	https://www.kaggle.com/sayangoswami/reddit-memes-dataset	Text Extraction: Optical Character Recognition (OCR) is used to extract text from meme images. Language Encoding: The extracted text is encoded using BERT representation for language, generating contextual word embeddings that are then averaged to form a sentence embedding. Visual Encoding: The visual information is encoded using a VGG-16 convolutional neural network trained on ImageNet. Activations from a hidden layer are used as feature vectors for the image. Multimodal Fusion: The encoded representations from language and vision are concatenated to form a multimodal representation. Classification: A multi-layer perceptron (MLP) is trained on the multimodal representation to predict the hate speech detection score.	Accuracy, Precision, Batch, Epoch	Multimodal: 0.823 Image: 0.804 Text: 0.750	Despite achieving some success, the task of hate speech detection in memes is far from being solved, given the high abstraction level of the messages contained in memes. The complexity of detecting hate speech in multimodal content poses challenges for automated systems.

3 Project Scope and Objectives

This research endeavors to address the complex and vital challenge of hate speech detection within memes, a context known for its intricate blend of humor, coded language, and visual elements. Hate speech can manifest in these digital cultural artifacts in subtle and hidden forms, often eluding traditional text-only analysis methods. To meet this challenge, our algorithm adopts a multimodal approach that

combines both text and image analyses. By concurrently considering linguistic and visual cues, this approach allows us to capture the nuanced interplay between textual and visual components within memes. The integration of sophisticated technologies, such as the BERT language model for text understanding and the VGG-16 Convolutional Neural Network for image analysis, enhances our algorithm's accuracy and effectiveness. Furthermore, Optical Character Recognition (OCR) is employed to extract textual content from images, contributing a vital dimension to the analysis process.[6]

Our overarching objective aligns with a larger mission—fostering inclusive and respectful online communities for all users. The development and implementation of this algorithm aim to contribute to the creation of safer online spaces, where users can engage without fear of encountering cyberbullying or hate speech. Recognizing the limitations and ethical considerations inherent in automated hate speech detection, our research places importance on gathering user feedback and ensuring algorithm adaptability to an ever-evolving online landscape. By targeting hate speech within memes, we aspire to uncover harmful content that may otherwise escape detection and promote online respect and inclusivity.[5]

4 Methodology

In our pursuit of developing an algorithm for the detection of cyberbullying and hate speech within memes, we adopt a comprehensive methodology to address the multifaceted challenges presented by this unique form of content. Central to our methodology is the utilization of purposefully curated datasets: the Hateful Memes Dataset by Facebook[15] and the Reddit Memes Dataset[14]. The Hateful Memes Dataset comprises over 10,000 multimodal examples, featuring both text and images, and is specifically tailored for understanding and identifying harmful content. To complement this, the Reddit Memes Dataset, consisting of 3,325 non-hate memes assumed to contain non-hateful content, serves as a crucial contrast and validation set. This dataset combination forms the foundation for comprehensive model training, validation, and ultimately, the effective detection of cyberbullying within memes.[6]

Our methodology is designed to explore the influence of distinct dimensions in enhancing the algorithm's performance. Initially, we train separate models using only the 768 dimensions of the BERT language model and the 4,096 dimensions from VGG-16. This allows us to evaluate the individual contributions of textual and visual components. Subsequently, we compare the accuracy of these individual models with the overarching model, which integrates both the 768 dimensions of BERT and the 4,096 dimensions of VGG-16, resulting in a 4,864-dimensional multimodal feature vector. This comparative analysis provides insights into the synergy achieved by our comprehensive model and demonstrates its potential for superior accuracy, surpassing the performance of the individual models.[6]

Our methodology is underpinned by functional requirements, including Optical Character Recognition (OCR) for text extraction, BERT encoding for text

understanding, and VGG-16 for image feature extraction. The fusion of these modalities enables the creation of a feature-rich multimodal representation, which is further processed by a multi-layer perceptron (MLP) with two hidden layers, each housing 100 neurons and the ReLU activation function. This approach aims to capture the intricate interplay between textual and visual content, thereby enhancing the algorithm's accuracy and effectiveness in detecting cyberbullying within memes.[6]

Additionally, we conduct in-depth analysis to explore the algorithm's sensitivity to nuanced language, as it is often used in memes that involve humor, sarcasm, or coded language. The integration of BERT's bidirectional processing and contextual comprehension, fine-tuned to align with meme-specific linguistic features, forms an essential component of our methodology.[6]

By conducting this thorough comparative analysis, our methodology aims to underscore the advantages of a multimodal approach, showcasing the algorithm's potential for superior performance. This research aligns with the broader mission of fostering an inclusive and respectful online community, ensuring user safety and well-being in the digital landscape.[6]

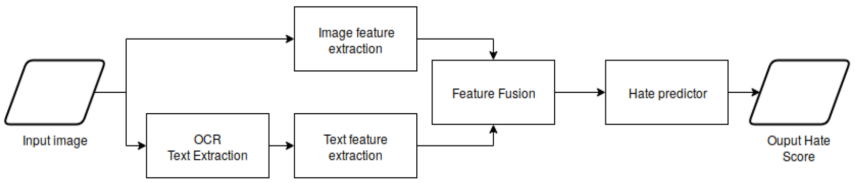

Fig. 2. The picture demonstrates our methodology and the overall design of our system. This block diagram illustrates how the image features and text features will be combined to determine if the given meme displays hateful tendencies or not.[16]

5 Implementation

Our methodology for the detection of cyberbullying within memes is underpinned by a sophisticated combination of tools and libraries. A crucial component of our approach is the application of Optical Character Recognition (OCR) technology, made possible through the utilization of the Tesseract 4.0.0 OCR engine, developed and maintained by Google. OCR serves as the initial step in our algorithm, allowing us to extract text from meme images. This capability is paramount, as memes frequently employ images alongside impactful text to convey messages and emotions. Tesseract 4.0.0's extensive reputation for accuracy and efficiency in text extraction equips us with a mature and robust technology to effectively handle the complexities presented by memes with varying fonts, styles, and sizes.[5]

5.1 **Usage of BERT:-**

In the quest to understand and identify potentially harmful content within memes, our algorithm takes advantage of the state-of-the-art language representation model, BERT (Bidirectional Encoder Representations from Transformers). Specifically, we employ the "bert-base-multilingual-cased" version of BERT, a variant designed to handle multiple languages while preserving case distinctions (uppercase and lowercase letters). This choice of BERT comes with 12 layers, each comprising 768 hidden dimensions and 12 attention heads. These attributes collectively enable the model to capture a broad spectrum of linguistic patterns, rendering it suitable for the analysis of memes that may contain content in diverse languages. BERT's strength lies in its bidirectional processing, which aids in understanding context—a vital aspect of detecting cyberbullying. This nuanced form of hate speech often relies on subtle phrasing, sarcasm, and coded language, making it challenging to identify using traditional keyword-based methods. BERT's capacity to comprehend context ensures that our algorithm can identify both overt and covert cyberbullying content. Fine-tuning BERT for meme-specific linguistic features is undertaken to align the model's understanding with these unique aspects of meme language.[5,6]

However, effective cyberbullying detection doesn't rely solely on textual analysis. Visual elements in memes also play a substantial role in conveying messages and detecting harmful content. To unlock these visual components, our algorithm leverages the VGG-16 Convolutional Neural Network (CNN), a well-established architecture renowned for its image analysis capabilities. VGG-16, with its 16 weight layers, excels in capturing intricate visual patterns. In our implementation, we employ a version of VGG-16 pretrained on the ImageNet dataset, a vast collection of labeled images across numerous categories. This pretraining empowers VGG-16 to recognize a wide spectrum of visual features that can be transferable to various image-related tasks, including cyberbullying detection within memes. Our algorithm extracts features from the last hidden layer of the VGG-16 model, representing abstract and high-level visual elements that the model has learned to identify within images. These features encompass basic elements such as edges and textures, but also extend to more complex structures like shapes, objects, and patterns.[5,6]

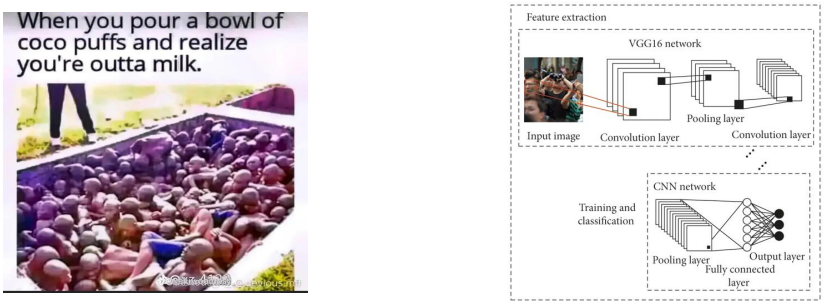

Fig. 3. The picture on the left demonstrates why only textual analysis would not work as that system would classify it as not cyberbullying whereas this is clearly cyberbullying with image context. The image on the right demonstrates image feature extraction using VGG16.[5,6]

5.2 **Usage of VGG-16s:-**

Recognizing that cyberbullying content in memes can be expressed through both textual and visual elements, our algorithm adopts a multimodal approach. This involves fusing the encoded text features from BERT with the extracted image features from VGG-16. The combination of these features results in a single, comprehensive feature vector that encapsulates both linguistic nuances and visual patterns. The concatenated feature vector is multidimensional, encompassing the sum of the dimensions of the individual encodings. Specifically, the text encoding from BERT contributes 768 dimensions, while the image encoding from VGG-16 adds 4,096 dimensions, resulting in a 4,864-dimensional multimodal feature vector. This multidimensional approach empowers our algorithm to capture complex interactions between textual and visual content, providing a more profound understanding of the meme's context and allowing for the identification of potentially harmful content that might be obscured when analyzing only one modality.[5,6]

5.3 **Combination of both BERT and VGG-16:-**

As the ultimate phase of our algorithm, the detection of hate speech within memes is carried out by a multi-layer perceptron (MLP). The MLP architecture, consisting of two hidden layers, each with 100 neurons and utilizing the Rectified Linear Unit (ReLU) activation function, excels in capturing intricate relationships between features. The choice of this architecture strikes a balance between complexity and computational efficiency. While more complex architectures with additional hidden layers and neurons might offer potential improvements in accuracy, they also increase the risk of overfitting and demand extensive computational resources. The final output of the MLP is a single neuron that produces a prediction score, reflecting the likelihood of the meme containing cyberbullying content. A higher output score corresponds to a higher probability of harmful content within the meme, and this score can be interpreted as a measure of the algorithm's confidence in its prediction.[5,6]

This comprehensive implementation platform, combining OCR, BERT, VGG-16, and the MLP, is carefully designed to address the multifaceted challenges presented by cyberbullying detection within memes. It fuses the strengths of textual and visual analyses to provide a holistic understanding of meme content and its potential for harm, aligning with the broader mission of fostering a safer and more inclusive online community.[5,6]

6 Activity Timeline Chart

In the initial phase of our research project, spanning from Week 1 to Week 2, we dedicated our efforts to extensive research and information gathering, primarily through a detailed literature review. Over a span of 4 days, we thoroughly analyzed research papers and gathered valuable insights related to the problem statement and existing works in the field. Simultaneously, we embarked on the crucial task of dataset acquisition and preparation. In the first part of this phase, which took 2 days,

we identified and collected pertinent datasets essential for our research objectives. Following this, we invested 3 days in meticulous data preprocessing, ensuring that the acquired data was suitably prepared for our model's analysis.

The subsequent phase, spanning from Week 2 to Week 3, primarily focused on model development. Over a course of 6 days, we rigorously implemented machine learning models using Optical Character Recognition (OCR), BERT for text analysis, VGG16 for image analysis, and a Multi-Layer Perceptron (MLP). Once the models were coded and ready, we dedicated 4 days to the critical process of model training and validation. This stage was pivotal in ensuring that the models could perform effectively and make accurate predictions.

Moving into Week 3, we transitioned to frontend development. Within a short duration of 2 days, we skillfully crafted the frontend interface to facilitate user interaction with the model. The subsequent step, taking 2 additional days, was dedicated to the seamless integration of the frontend with the machine learning model, marking the beginning of the testing phase.

The final stages of our project encompassed thorough integration and testing, with 2 days allocated for integrating the frontend with the machine learning model. Ultimately, the concluding phase, termed "Final Testing and Documentation," involved comprehensive testing to validate the model's functionality, ensuring its accuracy and reliability.

Throughout the project, a meticulous approach was maintained in line with our research objectives, with special emphasis on the critical areas of model development, data preparation, and seamless integration. The final deliverables are detailed in our conclusion and report, culminating our project and providing comprehensive insights for future reference.

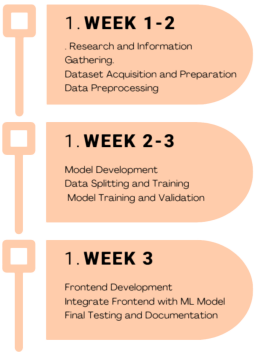

Fig. 4. The picture demonstrates our progress week wise in a flowchart.

7 Results

In this study, we trained the models on 3 sections of the same dataset. One of the models was trained on the text extracted from the image, the second model was trained on the image features and the third model was trained on both the image as well as text features.

Model 1 (Using only Text Features): Trained on extracted text from memes.
Accuracy: 77.94760320663452%
Model 2 (Image Features): Trained on image features extracted using VGG-16.
Accuracy: 84.97895167350769%
Model 3 (Combined Features): Trained on both image (VGG-16) and text (BERT) features.
Accuracy: 86.96432278633118%

Thus we can conclude that the model trained using both the image and the text features performed better than the model trained on only the text or the image features.

Results of Base Paper:-
Smth. Max Accuracy is the smoothened or converging accuracy of models when recorded multiple times
Max. Accuracy is the Maximum accuracy of Models recorded in any iteration.

Table 1: Accuracy results for the three configurations

Model	Max. Accuracy	Smth. Max. Accuracy
Multimodal	0.833	0.823
Image	0.830	0.804
Text	0.761	0.750

8 Conclusion and Future Scope

In this study, we have developed an innovative and comprehensive algorithm for the critical task of detecting cyberbullying within memes. Our approach uniquely

integrates both text and image analyses to ensure a more thorough and accurate identification of harmful content. The use of pre-trained models, such as BERT for text analysis and VGG-16 for image analysis, significantly enhances our algorithm's accuracy, allowing it to effectively address the multifaceted nature of memes where harmful content can be concealed behind humor, coded language, or visual elements.[5,6]

However, it is important to acknowledge several factors that impact our algorithm's performance. The quality, diversity, and size of the training dataset play a pivotal role in generalization, making it essential to continuously update and expand the dataset to maintain its real-world effectiveness. Furthermore, while BERT offers remarkable linguistic understanding, it may still struggle with nuanced elements such as sarcasm, irony, and cultural references within memes, necessitating ongoing model fine-tuning and adaptation.

Fig. 5. The picture on the left demonstrates a meme which BERT would have no problem in detecting text from. The picture on the right demonstrates a meme from which BERT might have trouble extracting text from.[5,6]

Our algorithm has demonstrated the significance of the visual modality in meme analysis. In essence, the context conveyed through images is paramount for accurate cyberbullying detection. However, it's crucial to remain vigilant against potential biases and ethical concerns, as automated hate speech detection algorithms can inadvertently flag certain content or misclassify innocuous materials.[5,6]

As we look to the future, our algorithm's scope extends beyond detection to adaptation and evolution. The dynamic nature of the online landscape and language usage demands regular updates and model refinements. Users' feedback and real-world experiences will be invaluable in shaping and improving the algorithm. Moreover, vigilance against potential misuse is vital, and we recognize the possibility of our system inadvertently aiding in the creation of hate speech content.

Consequently, continuous monitoring and ethical considerations remain central in our algorithm's ongoing development.

In conclusion, our multimodal algorithm harnesses the power of both language and visuals to effectively combat cyberbullying and hate speech within the realm of memes. By addressing the challenges posed by this unique form of content, we contribute to the creation of a safer and more inclusive online community, fostering an environment where users can interact without fear of harassment or harm. The future of our research is characterized by adaptability, continuous improvement, and the commitment to staying at the forefront of addressing evolving challenges in online communication.

9 Help Taken From Internet and Our Contribution

For this project, we took various research papers as reference to gain knowledge about the topics of cyberbullying, and how we can train a deep learning model to detect cyberbullying based on the text and image features extracted from the memes,

The dataset which we used for training and testing of out model are:
Kaggle Memes Dataset:
https://www.kaggle.com/datasets/sayangoswami/reddit-memes-dataset
Hateful Memes Dataset:
https://www.kaggle.com/datasets/parthplc/facebook-hateful-meme-dataset

We wrote python scripts for

1. segregating and splitting the large datasets,
2. generating metadata files for the datasets to use in the code while training the model
3. for extracting text from the memes using tesseract's ocr engine

We took inspiration from the base model used in the research paper :-

1. Hate Speech Detection in Pixels: https://arxiv.org/pdf/1910.02334.pdf

for our model.

The project which we used for creating an understanding of the topic of feature extraction from images, and using them to detect cyberbullying is:
https://github.com/imatge-upc/hate-speech-detection

The above repo contains some details about the model used in the above specified research paper.

The scripts used and the google colab notebook containing the code for training and testing the model can be accessed in this repo:-

https://github.com/rajat397/Meme-Based-CyberBullying-Detection

Overall, the contribution of our group in this project can be considered to be up to 70%, and 30% of the project work can be attributed to various internet sources and research papers we used for the training algorithms, example projects, and the annotated dataset to train and test the model.

References

2. Roushan Kumar Giri, Subhash Chandra Gupta, Umesh Kumar Gupta: An approach to detect offence in Memes using Natural Language Processing(NLP) and Deep learning: https://sci-hub.se/https://ieeexplore.ieee.org/document/9402406
3. Raul Gomez, Jaume Gibert, Lluis Gomez, Dimosthenis Karatzas Eurecat, Centre Tecnologic de Catalunya, Unitat de Tecnologies Audiovisuals, Barcelona, Spain, Computer Vision Center, Universitat Autonoma de Barcelona, Barcelona, Spain: Exploring Hate Speech Detection in Multimodal Publications: https://arxiv.org/pdf/1910.03814.pdf
4. Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, Paul Buitelaar: Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text: https://aclanthology.org/2020.trac-1.6.pdf
5. Haris Bin Zia, Ignacio Castro, Gareth Tyson: Racist or Sexist Meme? Classifying Memes beyond Hateful: https://aclanthology.org/2021.woah-1.23.pdf
6. Benet Oriol Sabat, Cristian Canton Ferrer, Xavier Giro-i-Nieto: Hate Speech in Pixels: Detection of Offensive Memes towards Automatic Moderation: https://arxiv.org/pdf/1910.02334.pdf
7. Lanyu Shang, Yang Zhang, Yuheng Zha, Yingxi Chen, Christina Youn, Dong Wang: AOMD: An Analogy-aware Approach to Offensive Meme Detection on Social Media: https://arxiv.org/pdf/2106.11229.pdf
8. Roy Ka-Wei Lee, Rui Cao, Ziqing Fan, Jing Jiang, Wen-Haw Chong: Disentangling Hate in Online Memes: https://dl.acm.org/doi/pdf/10.1145/3474085.3475625
9. Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, Tanmoy Chakraborty: DISARM: Detecting the Victims Targeted by Harmful Memes: https://aclanthology.org/2022.findings-naacl.118.pdf
10. https://github.com/imatge-upc/hate-speech-detection/blob/master/slides.pdf
11. https://en.wikipedia.org/wiki/Cyberbullying
12. https://en.wikipedia.org/wiki/Meme
13. https://knowyourmeme.com/photos/2385955-greentext-stories
14. https://preview.redd.it/ran-out-of-milk-v0-hgqfqxni8pya1.png?width=640&crop=smart&auto=webp&s=251f86a2175b2a210ae0f19d7ba14b15f193a2f0
15. https://www.kaggle.com/datasets/sayangoswami/reddit-memes-dataset
16. https://www.kaggle.com/datasets/parthplc/facebook-hateful-meme-dataset