International Research Journal of Education and Technology
Peer Reviewed Journal
ISSN 2581-7795

Hateful Meme Detection Using Deep Learning

S. Kishorebalaji[1], A. Pradeep[2], R. Ramesh Kannan[3],A. Shenbagapriya[4]

[1]S. Kishorebalaji, CSE, Sri Ramakrishna Institute of Technology, Coimbatore, Tamil Nadu, India.
[2]A. Pradeep, CSE, Sri Ramakrishna Institute of Technology, Coimbatore, Tamil Nadu, India.[3]R. Rameshkannan, CSE, Sri Ramakrishna Institute of Technology, Coimbatore, Tamil Nadu, India. [4]Ms. A. Shenbagapriya, Assistant Professor, Department of CSE, Sri Ramakrishna Institute of Technology, Coimbatore, India.

Abstract:

Detecting hateful memes across different types of media, like images and audio, is becoming increasingly important in today's digital world, especially given their widespread impact on social media. To tackle this issue, we're developing a web-based app that uses deep learning models—such as CNNs and RNNs—along with multimodal techniques like CLIP to spot hateful content. The goal is to reduce the negative effects of harmful memes in our online communities. This project is designed with ethical standards in mind, ensuring that digital content is shared responsibly. In the end, it aims to create a safer, more positive online space for everyone

1. INTRODUCTION

Social media platforms have become a primary means of communication, but they also facilitate the spread of harmful content, including hateful memes. As digital interactions increase, the need for automated content moderation becomes crucial to maintain a safe online environment. AI-powered hateful meme detection systems enhance online safety by analyzing text and images, identifying offensive content, and reducing the spread of hate speech.This project aims to develop a deep learning-based hateful meme detection system that leverages multimodal analysis. By combining image classification (CNNs) and speech/text analysis (RNNs/ASR), the system accurately detects and flags hateful content. This approach enhances content moderation efforts, ensuring a healthier and more inclusive digital space.

1.1. General Introduction

The Social media has become a dominant medium for communication, but it also serves as a breeding ground for hateful content, particularly through memes that combine offensive imagery and text. Traditional content moderation methods struggle to detect these harmful memes, as they often rely on manual review or outdated detection systems that fail to understand multimodal content. This leads to the unchecked spread of hate speech, fostering toxicity in online communities.This project introduces an AI-powered hateful meme detection system that utilizes deep learning models for image and text analysis. By leveraging CNNs for image processing and RNNs for speech/text analysis, the system enhances content moderation by accurately identifying and flagging hateful memes. By automating hate speech detection, this project aims to create a safer online environment, reducing the spread of offensive content while promoting inclusive digital interactions.

1.2. Problem Statement

In With the rise of social media, hateful content—especially in the form of memes—has become a growing concern. These memes often contain harmful imagery, offensive text, or disguised hate speech, making them difficult to detect with traditional moderation techniques. Existing content filtering systems primarily rely on keyword detection or manual review, which are inefficient, inconsistent, and fail to interpret the multimodal nature of memes. As a result, harmful content continues to spread, negatively impacting online communities.
This project introduces an AI-powered hateful meme detection system that leverages deep learning models to analyze both visual and textual content. By utilizing CNNs for image classification and RNNs for speech/text processing, the system accurately identifies hateful memes, enhancing content moderation. This approach reduces the burden on human moderators, improves detection efficiency, and helps create a safer digital environment by preventing the spread of offensive content.

2. LITERATURE SURVEY

1. Schmidt et al., (2017) expressed a comprehensive evaluation of hate speech detection strategies, particularly within social media structures. the author classifies the prevailing detection techniques into three fundamental categories: rule-based totally techniques, gadget studying strategies, and deep mastering strategies. Rule-based totally methods depend upon predefined lexicons and heuristics, which limits their effectiveness due to their incapability to adapt to the changing nature of hate speech. however, gadget learning strategies, consisting of help Vector Machines (SVM) and Logistic Regression, allow for the introduction of fashions that research from labeled datasets. Deep learning techniques, mainly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), are emphasized for his or her functionality to routinely extract features from text, which enhances detection accuracy. The evaluate additionally highlights several great demanding situations, which includes the evolving nature of language, the paradox of context, and the need for culturally conscious detection systems. the author shows that destiny research should consciousness on multimodal techniques that integrate visible and textual records to enhance the accuracy of hateful.

2. Aguirre et al., (2019) classified current techniques in keeping with the fusion techniques they adopt, which includes early, late, and hybrid fusion techniques. Early fusion combines capabilities from distinct belief modalities, even as early fusion combines itself from each change. Hybrid fusion gives methods to leverage the strengths of each. This paper discusses various modeling strategies, which include CNNs for picture analysis and RNNs for phrase processing, and their application to dispensed reasoning duties. This evaluate highlights the importance of the context and interplay of various fashions in enhancing the effectiveness of mental theories. The authors suggest that a specific technique to hate speech may be useful. cope with the challenge of identifying poor content on social media by means of combining visual content with information to be explored.

3. Ferreira et al., (2020) investigated methodologies and challenges associated with detecting hate speech in photograph content material, consisting of memes and social media posts. The authors present an outline of traditional photograph processing techniques, including edge detection and shade histograms, and evaluation these with current deep getting to know methods, such as CNNs and Generative opposed Networks (GANs). The paper highlights the significance of context in visual content, noting that symbols and imagery can convey hate speech even

4. Rehmani et al., (2020) focused at the software of deep studying techniques in hate speech detection, highlighting numerous architectures inclusive of CNNs, RNNs, and Transformers. The authors discover the benefits of deep getting to know, such as its capability to automatically learn functions from raw information without great manual feature engineering. The overview discusses particular packages of CNNs for picture primarily based hate speech detection and RNNs for text evaluation, noting their effectiveness in managing sequential facts. The paper additionally addresses commonplace demanding situations in deep gaining knowledge of, consisting of overfitting and the want for big annotated datasets. The authors stress the importance of integrating both textual and visual statistics in multimodal detection systems to enhance performance and robustness. They provide insights into destiny guidelines for studies, which include the exploration of unsupervised and semi-supervised mastering techniques to lessen dependency on classified facts.when the accompanying text is beign. The authors analyze several publicly available datasets designed for training and evaluating hate speech detection fashions, discussing their strengths and weaknesses. Key challenges identified consist of the want for context-conscious models that can recognize cultural references and the consequences of ambiguity in visual content. The authors call for the development of hybrid fashions that may combine textual and visual evaluation for greater accurate detection of hate speech in pix.

5. Yang et al., (2021) cited the role of contrastive mastering in improving multimodal sentiment analysis systems. The authors provide an explanation for how contrastive reading works via education models to distinguish between comparable and extraordinary statistics elements, efficaciously enhancing characteristic representation in the course of modalities.The paper critiques numerous multimodal datasets used for sentiment evaluation, in addition to contemporary improvements in version architectures that comprise contrastive reading techniques. The authors spotlight the functionality for those techniques to beautify hate speech detection through aligning seen and textual talents, therefore presenting a extra entire statistics of the context in which hate speech takes place. The evaluation discusses the consequences of multimodal getting to know for actual-worldwide

programs, which incorporates the importance of ethical concerns in deploying such systems.

6. Bechara et al., (2021) provided a comprehensive review that synthesizes research on hate speech, offering a clear definition and theoretical framework for understanding this complex issue. The authors examine various detection methodologies, categorizing them into three types: text-based, image-based, and multimodal approaches. They underscore the significance of context in interpreting hate speech, noting that the same phrase can have different meanings depending on the surrounding content. The paper addresses the limitations of current methodologies, particularly in their ability to grasp cultural nuances and adapt to evolving language use. The authors advocate for interdisciplinary approaches that integrate insights from linguistics, sociology, and computer science to improve detection capabilities. They conclude by emphasizing the need for ongoing research into context-aware detection systems that can adjust to the dynamic nature of hate speech.

7. Azad et al., (2021) presented an in depth assessment of the various techniques hired in hate speech detection, categorizing them into traditional system gaining knowledge of methods, deep getting to know techniques and hybrid fashions. The authors discuss the common challenges confronted inside the area, which includes dataset imbalance, that may cause biased model predictions, and the evolving nature of hate speech that complicates detection efforts. The assessment additionally addresses the issue of characteristic representation, emphasizing the want for fashions that can capture diffused nuances in language and context. The authors endorse for using multimodal records, along with pix and videos, to decorate detection talents. they also talk moral concerns, along with the ability for over-censorship and the significance of retaining free speech even as effectively addressing hate speech on line. The paper concludes with 8 suggestions for future research, inclusive of the development of more strong datasets and models that can adapt to converting linguistic patterns in hate speech. Datasets used for sentiment analysis, as nicely as recent advancements in model architectures that contain contrastive gaining knowledge of techniques. The authors highlight the potential for those approaches to improve hate speech detection through aligning visual and textual functions, for that reason supplying a extra complete expertise of the context in which hate speech occurs. The evaluation discusses the results of multimodal getting to know for real-international applications, inclusive of the significance of moral concerns in deploying such systems.

8. Al-Sharif et al., (2022) reviewed synthesizes modern-day studies on text-based hate speech detection, specializing in severa gadget reading techniques. The authors categorize the techniques based on characteristic extraction techniques, which incorporates bag-of-words, TF-IDF, and phrase embeddings, analyzing their effectiveness in taking pictures the semantics of hate speech. They examine unique classifiers, which includes desire wooden, SVMs, and ensemble strategies, discussing their strengths and weaknesses in diverse contexts. The paper additionally highlights the importance of preprocessing steps, in conjunction with stemming and lemmatization, in improving version overall performance. The authors call for greater comprehensive evaluations of hate speech detection systems, suggesting the combination of multimodal data to enhance detection accuracy and robustness.

9. Ali et al., (2023) addressed the multifaceted challenges of detecting hate speech across distinctive media formats, noting both technical and moral issues. On the technical side, they discuss obstacles like statistics shortage, which hinders the creation of robust models, and the problem of model interpretability, which influences transparency. Ethical demanding situations are similarly giant, in particular concerns approximately censorship and maintaining freedom of speech in detection systems. The authors have a look at current advancements within the area, including records augmentation strategies that amplify limited datasets and switch getting to know strategies that enhance model adaptability. Additionally, they highlight gaps in current studies and emphasize the want for inclusive datasets that extra appropriately mirror diverse cultural views, ensuring that dislike speech detection equipment are honest and effective throughout diverse contexts. The paper emphasizes the significance of collaboration amongst researchers, policymakers, and social media systems to construct comprehensive answers that stability technical effectiveness with ethical duty. Ultimately, the authors suggest for interdisciplinary techniques that combine technical, social, and ethical viewpoints, arguing that those are essential for addressing the complicated, multimodal

10. Aggarwal (2023) proposed that AI can significantly elevate the online grocery shopping experience. By analyzing customer behaviors and preferences, AI algorithms create tailored product recommendations, helping customers find new items that align with their tastes. The study also highlights the role of AI-driven chatbots in providing instant customer support and enhancing satisfaction by addressing queries quickly and efficiently. For retailers, this technology not only boosts customer engagement but also drives sales and streamlines operations. The use of machine learning allows retailers to refine recommendations continuously, ensuring they remain relevant to evolving customer preferences.

3. EXISTING METHODOLOGY

Current content moderation systems primarily rely on manual review, keyword-based filtering, and rule-based algorithms to detect hateful content. While these methods provide some level of moderation, they often struggle with detecting hateful memes, as offensive content is embedded in both images and text. Keyword-based filtering fails to recognize hidden meanings, sarcasm, or implicit hate speech, while manual moderation is time-consuming, inconsistent, and prone to human bias. Additionally, most traditional systems lack multimodal analysis, making it difficult to interpret the relationship between text and images. To address these limitations, our project introduces an AI-powered hateful meme detection system that leverages CNNs for image classification and RNNs for text/audio analysis. By integrating deep learning-based multimodal analysis, the system ensures more accurate and automated detection, significantly improving content moderation and contributing to a safer online environment.

3.1. DISADVANTAGES

1. **Limited Detection Accuracy –** Traditional content moderation systems, such as keyword filtering and rule-based approaches, often fail to detect hidden hate speech, sarcasm, and implicit offensive content, reducing overall effectiveness.
2. **Lack of Multimodal Analysis** – Existing methods primarily analyze text or images separately, making it difficult to interpret memes where hate speech is embedded in both modalities.
3. **High Dependence on Manual Moderation –** Many platforms still rely on human reviewers, which is time-consuming, inconsistent, and prone to bias, leading to delays in content moderation.
4. **Inability to Adapt to New Hate Speech Trends –** Static rule-based systems struggle to keep up with evolving offensive language,and emerging meme formats, reducing their long-term effectiveness.
5. **Scalability Issues –** As social media platforms generate vast amounts of content daily, existing moderation techniques struggle to process data efficiently, requiring significant resources to scale.

4. PROPOSED METHODOLOGY

The proposed system is an hateful meme detection model designed to enhance content moderation on social media platforms. Its primary goal is to accurately detect and classify hateful memes by analyzing both visual and textual elements using deep learning-based multimodal analysis.

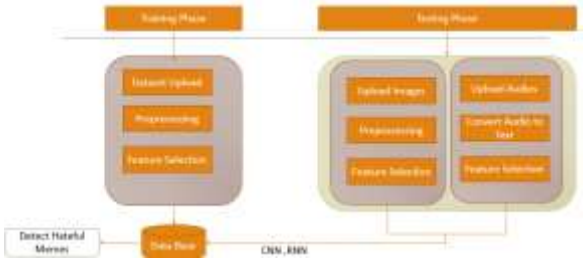
By leveraging advanced AI techniques, the system integrates Convolutional Neural Networks (CNNs) for image processing and Recurrent Neural Networks (RNNs) for text and speech analysis. This allows for a comprehensive understanding of offensive content, even when hate speech is embedded within memes through subtle visual cues, implicit text, or speech tone variations. The system further refines its accuracy by utilizing Automatic Speech Recognition (ASR) for extracting text from audio-based memes and applying Natural Language Processing (NLP) for sentiment analysis.

Additionally, the system ensures efficient real-time detection by continuously learning from new trends and evolving hate speech patterns. By automating the identification and classification of hateful memes, this approach significantly improves content moderation, reduces the burden on human reviewers, and contributes to a safer and more inclusive digital space.

4.1. ADVANTAGES

1. **Enhanced Accuracy –** The system uses CNNs for image analysis and RNNs for text/audio processing, ensuring precise detection of hateful memes.
2. **Real-Time & Scalable Moderation –** AI-powered automation enables fast and efficient content filtering, reducing reliance on manual moderation.
3. **Adaptive Learning –** The system continuously updates itself to detect evolving hate speech trends, improving long-term effectiveness.
4. **Accurate Hateful Content Detection** – The system effectively identifies hateful memes by analyzing both visual and textual elements, improving detection accuracy compared to traditional methods.
5. **Automated and Scalable Moderation** – Unlike manual review methods, the AI-powered system can process large volumes of content in real time, making it highly scalable for social media platforms

4.2. BLOCK DIAGRAM



5. RESULTS

5.1. HOME PAGE

The homepage features a clean and user-friendly interface, designed to provide a seamless experience for users. A navigation bar at the top allows access to key sections such as Home, About, Detection, and Contact Us, along with a Sign-In button for user authentication. At the center of the page, users can upload or input memes for hateful content analysis, making detection quick and accessible. The background incorporates modern and minimalistic design elements, ensuring a professional and engaging look. This homepage serves as the entry point for users to explore the AI-powered hateful meme detection system, enabling efficient and accurate content moderation with ease.

5.2. LOGIN PAGE

The login page is designed to provide a secure and seamless authentication process for users accessing the hateful meme detection system. Users can sign in using their credentials or opt for third-party authentication via Google or other social platforms. The system securely manages user data using a robust database, ensuring efficient handling of user profiles and login sessions.Once logged in, users can upload memes for analysis, view detection history, and access moderation tools. The user-friendly interface includes input fields for email and password, a forgot password option, and a sign-up link for new users, ensuring an intuitive experience. The secure authentication process helps maintain the platform's integrity, supporting a safe and controlled environment for AI-driven content moderation.

5.3. IMAGE PREDICTION

The Prediction Page is a key component of the Hateful Meme Detection system, allowing users to upload meme images and instantly receive a classification result. The interface features a simple file upload section and a “Detect Hateful Content” button, making it easy for users to interact with the system. Once an image is uploaded, the AI model processes the content and displays a result indicating whether the meme contains hateful elements, such as the example result: “Is Hateful: Yes.”
This section plays a crucial role in ensuring efficient and real-time content moderation. By providing quick and accurate results, it empowers users to identify and address harmful content before it spreads. The clear layout and immediate feedback enhance the overall usability of the platform, promoting a safer and more responsible digital environment.

5.4. AUDIO DETECTION

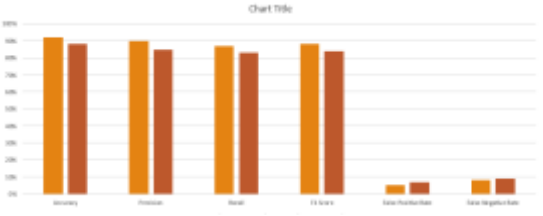
The audio meme detection system shown in the screenshots demonstrates how uploaded audio files are analyzed to determine whether they contain hateful content. When a user uploads an audio file, the system first transcribes the speech into text using speech-to-text technology. This transcribed text is then analyzed by an underlying hate detection model, which determines if the content is hateful or not based on certain keywords, tone, context, and patterns it has learned during training. In the first example, the audio file named "harvard.wav" was uploaded and transcribed into a sentence that, although disjointed and seemingly random, was flagged

as hateful. This suggests that the model may have detected certain combinations of words or sentiments that are commonly associated with toxic or inappropriate language. On the other hand, in the second example, the audio file "The quick brown fox.wav" was transcribed as the well-known pangram "the quick brown fox jumps over the lazy dog," and it was rightly identified as non-hateful. This clear and neutral sentence serves as a good example of how the system can distinguish between harmful and harmless content. Overall, this feature is highly useful for platforms or applications aiming to filter user-generated audio content, ensuring that users are alerted if any uploaded voice data contains speech that could be considered offensive or harmful before it is shared or published online. The user-friendly interface, with its straightforward upload function and instant feedback, makes it easy for anyone to understand the nature of the content they are working with, promoting responsible digital interactions.

5.5 PERFORMANCE EVALUATION

The performance evaluation of the hate speech detection models is illustrated through several key metrics, including Accuracy, Precision, Recall, F1 Score, False Positive Rate, and False Negative Rate. As shown in the bar chart, both models demonstrate strong performance with high values across Accuracy, Precision, Recall, and F1 Score—ranging above 85%, indicating that the models are effective in correctly identifying hateful content. Slight variations exist between the models, with one showing a marginal advantage in terms of overall classification performance, particularly in Precision and F1 Score.

In contrast, the False Positive Rate and False Negative Rate remain relatively low for both models, reflecting their reliability in avoiding incorrect classifications. These low error rates are critical for maintaining the credibility of a hate detection system, ensuring that non-hateful content isn't wrongly flagged and that harmful content isn't overlooked. Overall, the metrics suggest that both models are well-optimized and capable of performing robust hate speech classification with minimal misclassification.



6. CONCLUSION

This project introduces an AI-powered e-commerce recommendation system, designed to enhance the online shopping experience through intelligent and personalized suggestions. By analyzing user interactions, past selections, and product data, the system delivers tailored recommendations, helping users discover relevant items effortlessly. Key features include an intelligent recipe recommendation system that suggests recipes along with required ingredients, step-by-step cooking assistance with interactive guidance, real-time stock management to ensure product availability, and a comprehensive admin dashboard for seamless order and inventory tracking. Additionally, secure payment processing and advanced product categorization streamline the shopping experience. By addressing challenges such as choice overload and user engagement, the system fosters customer satisfaction and loyalty. This project highlights the transformative role of AI in e-commerce, optimizing decision-making, improving user experience, and enhancing operational efficiency for both customers and providers.

7. REFERENCE

[1] C. W. Alorainy, P. Burnap, H. Liu, and M. L. Williams, '"The enemy among us': Detecting cyber hate speech with threats-based othering language embeddings," ACM Trans. Web, vol. 13, no. 3, pp. 1–26, Aug. 2019

[2] N. S. Mullah and W. M. N. W. Zainon, "Advances in machine learning algorithms for hate speech detection in social media: A review," IEEE Access, vol. 9, pp. 88364–88376, 2021

[3] T. X. Moy, M. Raheem, and R. Logeswaran, ‘‘Hate speech detection in English and non-English languages: A review of techniques and challenges,’’ Webology, vol. 18, no. 5, pp. 929–938, Oct. 2021.

[4] J. Badour and J. A. Brown, "Hateful Memes Classification using Machine Learning," IEEE Symposium Series on Computational Intelligence (SSCI), Orlando, FL, USA, pp. 1-8, 2021

[5] Y. Zhou, Z. Chen and H. Yang, "Multimodal Learning For Hateful Memes Detection," 2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Shenzhen, China, pp. 1-6, 2021.

[6] Zhang, Y.; Yang, Q. A survey on multi-task learning. IEEE Trans. Knowl. Data Eng, 34, pp. 5586–5609, 2021.

[7] Hiril,.E,W.Pamungkas,F.Benamara,V.Moriceau, and V. Patti, ‘‘Emotionally informed hate speech detection: A multi-target perspective,’’ vol. 14, no. 1, pp. 322–352, Jan. 2022.

[8] R. T. Mutanga, N. Naicker, and O. O. Olugbara, ‘‘Detecting hate speech on Twitter network using ensemble machine learning,’’ Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 3, pp. 331–339, 2022,

[9] Ma, Z.; Yao, S.; Wu, L.; Gao, S.; Zhang, Y. Hateful Memes Detection Based on Multi Task Learning. Mathematics, vol. 10, pp. 4525, 2022.

[10] Z. Mansur, N. Omar and S. Tiun, "Twitter Hate Speech Detection: A Systematic Review of Methods, Taxonomy Analysis, Challenges, and Opportunities,"