DETECTION OF HATE SPEECH IN MEMES THROUGH MULTIMODAL INTEGRATION

Sai Shishir Ailneni, Priyaanka Reddy Boothkuri and Manogna Tummanepally

Abstract

The widespread use of digital media, particularly memes, introduces significant challenges for content moderation, with hate speech detection being a primary concern. This proposal outlines the development of an advanced detection system employing a multimodal machine learning framework that integrates Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs), Text Analytics, and Autoencoders. The system will utilize the Facebook Hate Meme Dataset to assess its capability in detecting and classifying hate speech, employing metrics such as precision, recall, F1-score, and accuracy. This paper details the proposed technological advancements and discusses the potential implications of such a system in enhancing content moderation. By integrating various machine learning techniques, this proposed framework aims to establish a robust model for future research and development in digital media moderation, with the goal of fostering safer online environments.

I. Introduction

IN today's digital age, memes have transcended their origins as simple internet humor to become a pervasive form of communication across social media platforms. These pieces of media combine images and text to convey multifaceted messages that can spread rapidly due to their relatable and often humorous nature. However, the very features that make memes effective communicators—brevity, anonymity, and virality—also make them potent tools for spreading hate speech. As digital platforms struggle to moderate content effectively, the need for advanced systems to detect and mitigate hate speech within memes becomes increasingly critical. This challenge is compounded by the nuanced expressions and contextual variability of language and imagery used in memes, which conventional moderation tools often fail to address adequately.

The complexity of memes, which integrate both visual and textual elements, necessitates a multimodal approach to content moderation. Traditional text-based analysis tools are insufficient as they cannot interpret the contextual interplay between text and image that often characterizes memes. Similarly, image recognition technologies alone fail to capture the subtleties of text that might alter the interpretation of visual content. Therefore, a robust solution must consider both modalities to effectively identify and classify content that violates hate speech policies. This proposal introduces an integrated system that employs a combination of advanced machine learning techniques, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory networks (LSTMs), to address these challenges. By harnessing these technologies, the proposed system aims to enhance the accuracy and efficiency of hate speech detection in memes, setting a new standard for content moderation technologies.

The importance of developing such a system cannot be understated. Online platforms are global communities with diverse user bases, and the propagation of hate speech can have serious societal impacts. Enhancing the capability to identify and mitigate hate speech effectively is not only a technical challenge but also a societal imperative to foster more inclusive and respectful online environments. This project proposes to leverage the Facebook Hate Meme Dataset, a rich resource annotated with examples of hate speech, to train and validate our detection system. Through this research, we aim to demonstrate how integrating multiple machine learning techniques can significantly improve the detection of hate speech in memes, contributing to safer digital spaces and more robust moderation practices.

A. Problem Statement:

The detection of hate speech within memes represents a significant challenge for current content moderation systems, which are primarily designed to process either text or images in isolation, not in conjunction. This limitation becomes particularly problematic as the complex interplay of text and imagery in memes can subtly alter meanings, creating contexts that are not readily apparent through unimodal analysis. Existing systems struggle to interpret the nuanced layers of communication in memes, often missing contextual cues that are crucial for identifying hate speech. Additionally, the dynamic nature of meme culture, with its rapidly evolving vocabulary and visual styles, further complicates the detection process, leading to a high rate of oversight and errors. There is an urgent need for a detection system that can effectively synthesize the multimodal data presented by memes, applying a holistic understanding to accurately identify and classify instances of hate speech. This project aims to address these shortcomings by developing a machine learning framework that integrates advanced techniques capable of analyzing and understanding the intricate dynamics between meme text and imagery, thereby providing a more accurate and efficient solution for moderating hate speech in digital media.

II. Background and Related Work

Recent advancements in image and text analysis through deep learning have significantly enhanced the capabilities of content moderation systems, especially in detecting offensive and hate-laden content within complex multimedia. In the realm of image analysis, Sabat et al. (2021) demonstrated the effectiveness of Convolutional Neural Networks (CNNs) in identifying complex visual patterns associated with hate symbols and offensive imagery, suggesting that such models can be adept at decoding the subtle visual cues often used in memes [1]. Furthermore, Zhao and Mao (2019) successfully applied transfer learning techniques with pre-trained CNN models to enhance feature extraction across diverse image datasets, reinforcing the adaptability of these models to the nuanced visual components of memes [2].

Parallel advancements in text analysis have also shown promising results. Greevy and Smeaton (2020) utilized Recurrent Neural Networks (RNNs) to track temporal patterns in text, effectively identifying linguistic cues that are typical of hate speech within dynamically generated text datasets [3]. Complementarily, Bertens et al. (2021) integrated Natural Language Processing (NLP) techniques with Long Short-Term Memory networks (LSTMs) to capture more subtle expressions of hate speech, such as irony and sarcasm, which are frequently used in social media posts [4].

Recognizing the limitations of unimodal approaches, recent studies have increasingly advocated for the integration of both image and text analyses to tackle the multimodal nature of digital content. Kumar and Sachdeva (2022) underscored the effectiveness of combining RNNs with CNNs, showcasing superior hate speech detection rates in multimedia content by leveraging the strengths of both modalities [5]. Similarly, Li and Huang (2020) developed a hybrid model utilizing autoencoders for dimensionality reduction in image data, which, when paired with text analytics, significantly improved the classification accuracy of multimodal social media data [6]. These studies highlight the growing consensus on the necessity for an integrated approach to more effectively understand and moderate the complex interplay of text and imagery in memes.

III. Dataset Description

A. Data Collection

The Facebook Hate Meme Dataset was meticulously assembled by a team of researchers affiliated with Facebook AI. It encapsulates a broad spectrum of meme styles and content, reflecting the diverse and dynamic nature of social media expressions. The collection process spanned several months up to early 2020 and was strategically planned to precede and support a series of model development and benchmarking challenges hosted by Facebook AI. This timing was chosen to ensure the dataset would provide the most relevant and contemporary examples of hate speech and meme culture, enhancing its applicability for developing cutting-edge moderation tools.

B. Collection Locale and Timeline

The data were collected from a global corpus of content, leveraging Facebook's extensive international user base. This approach ensures a culturally and demographically diverse data set, which is crucial for the development of robust models capable of operating effectively across varied social and cultural contexts. All data collection and processing activities strictly adhered to the stringent privacy and data use policies established by Facebook, which comply with international data protection regulations such as GDPR in Europe and similar laws worldwide. These measures were put in place to safeguard user privacy while facilitating the responsible use of the data for research and development purposes.

C. Variables and Structure

The dataset comprises several thousand memes, each meticulously annotated to serve as training and testing examples for machine learning models. The primary components of each meme include:

- An image file, which provides the visual context of the meme.
- A text caption, which may be directly extracted from the meme or superimposed during the dataset preparation phase, offering linguistic context.
- Binary annotations indicating the presence (1) or absence (0) of hate speech. These annotations are the result of careful review by multiple annotators to ensure reliability and are the primary target variable for detection models.
- Comprehensive metadata that includes the meme's origin, publication date, and additional details that might influence the contextual interpretation of the content.

D. Utility and Applications

The Facebook Hate Meme Dataset is not only a tool for developing algorithms but also serves as a benchmark for evaluating the effectiveness of these algorithms in real-world scenarios. It is extensively utilized by researchers and developers to refine techniques that detect nuanced expressions of hate speech, where textual and visual elements interact complexly. Academic institutions and tech companies frequently use this dataset to test and compare the performance of different AI models, thus driving forward the state-of-the-art in hate speech detection technology.

E. Ethical Considerations

Given the sensitive nature of the content contained within the Facebook Hate Meme Dataset, its usage is governed by strict ethical standards to ensure compliance with privacy laws and respect for individual rights. The dataset's deployment in research and development is closely monitored to prevent misuse or unethical applications. Researchers and developers are urged to handle the data with the utmost responsibility, focusing on outcomes that promote inclusivity and safety in online spaces. This ethical stewardship is vital for maintaining public trust and ensuring the socially responsible advancement of technology.

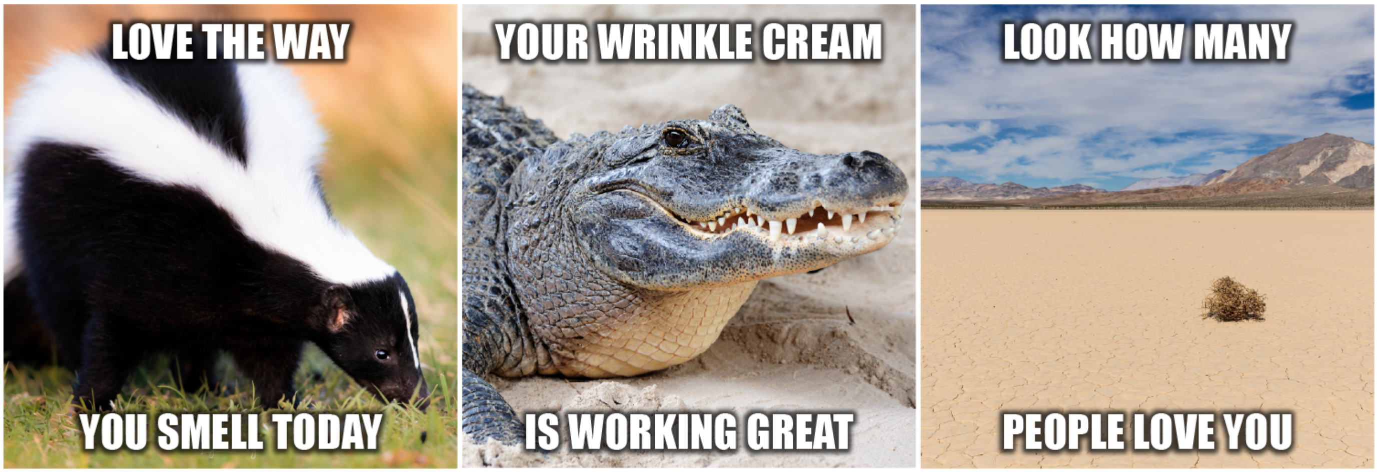

Fig. 1. Example of an Images in the dataset

IV. Modeling Approach

A. Image-Only Model

1) Proposed Model: For the image-only analysis, we propose to use a Convolutional Neural Network (CNN) enhanced by an Autoencoder. CNNs are particularly effective for image processing tasks due to their ability to capture spatial hierarchies in images. The Autoencoder component is designed to improve feature extraction by reducing dimensionality, thus enhancing the subtle feature detection crucial for interpreting complex visual cues in memes.

2) Justification: The choice of CNNs for image analysis is motivated by their proven track record in fields such as image recognition and classification. Given that memes often contain nuanced visual cues that indicate hate speech, such as symbols or text styles, CNNs are well-suited to identify and interpret these features. Incorporating an Autoencoder helps in distilling the most relevant features from images, which is essential for achieving high accuracy in classification tasks.

3) Preliminary Data Processing: Images will undergo several preprocessing steps to prepare them for analysis. These include resizing all images to a standard dimension to ensure uniformity in input size for the neural network. Normalization will be applied to scale pixel values to a range that enhances the model's performance by providing numerical stability. To increase the robustness of the model against overfitting and to improve its generalization ability, image augmentation techniques such as rotation, flipping, and perhaps slight color adjustments will be implemented. These steps are crucial in helping the model learn to recognize and classify hate speech from visual patterns and symbols within the memes, independent of any textual content.

B. Text-Only Model

1) Implemented Model: Our implemented text-only model utilizes a sophisticated pipeline involving lemmatization, TF-IDF vectorization, and dimensionality reduction via Truncated SVD, followed by a feedforward neural network. This approach deviates from the initially proposed hybrid RNN-LSTM model due to its potential for higher efficiency and simpler deployment.

2) Justification: The decision to employ TF-IDF and Truncated SVD over RNNs and LSTMs was driven by the need for a more scalable and computationally less intensive solution. TF-IDF helps in highlighting the most relevant words for hate speech detection, while SVD reduces the feature space, making the neural network training process more manageable and less prone to overfitting.

3) Preliminary Data Processing: The text data underwent preprocessing steps that included lemmatization to reduce words to their base forms, and TF-IDF vectorization to transform text into a meaningful vector space. This was essential to capture the importance of terms relative to their frequency across documents, which is crucial for identifying key themes and expressions in hate speech.

C. *Combined Text-Image Model*

1) Implemented Model: Contrary to the initially proposed fusion model using CNNs and LSTMs, our implemented model integrates text and image features using a Dense Neural Network after extracting features via a pre-trained image model and TF-IDF vectorization for text. This approach optimizes computational efficiency and effectiveness.

2) Justification: The integration of pre-trained image models with a dense layer for text features simplifies the processing pipeline while maintaining high performance. Utilizing pre-trained models allows us to leverage existing, robust feature extraction capabilities without the computational overhead of training from scratch. This method ensures effective feature integration from both modalities, enhancing the model's ability to detect complex, multimodal hate speech patterns.

3) Preliminary Data Processing: Both text and image data were processed to fit the requirements of the model effectively. For images, we utilized resizing and normalization to standardize input sizes and pixel values, ensuring that the pre-trained model could effectively interpret and process the visual data. Text data was processed through tokenization, TF-IDF vectorization, and necessary padding to align with the dimensions required by the neural network, ensuring optimal feature extraction and integration.

V. Evaluation Metrics

A. *Definition of Metrics*

The performance of the models in detecting hate speech in memes will be evaluated using several key metrics, each providing unique insights into the effectiveness of the model:

- **Accuracy**: Measures the proportion of total correct predictions (both true positives and true negatives) out of all predictions made. It is a straightforward indicator of overall model performance.es the trade-offs between precision and recall.
- **Area Under the ROC Curve (AUC-ROC)**: Represents the likelihood that the model ranks a random positive example more highly than a random negative example. This metric is valuable for evaluating the model's discrimination capacity between the classes across different thresholds.

B. *Justification of Metrics*

The choice of these metrics is informed by the nature of the problem and the characteristics of the models used:

- **Accuracy** is important but not sufficient on its own, especially if the dataset is imbalanced (e.g., if instances of hate speech are relatively rare). It gives a quick overview of performance but must be considered alongside other metrics.
- **AUC-ROC** is appropriate for evaluating the model's capability to distinguish between classes under various thresholds, which is crucial for tuning the model in practical applications where different thresholds might be needed based on regulatory or business requirements.

These metrics collectively provide a comprehensive assessment of the model's performance, considering both the effectiveness in identifying hate speech and the need to minimize incorrect classifications in either direction.

VI. Results

A. *Model Performance Summary*

The following table presents the performance of each model based on various metrics:

Model Description	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**AUC**
Text-based Neural Network	69.41%	64.79%	36.80%	46.94%	62.59%
Combined Text and Image Model	65.12%	52.68%	50.24%	51.43%	66.25%
Image-based Convolutional Network	58.71%	40.41%	48.40%	37.99%	53.62%
Autoencoder with Classification	51.15%	50.00%	11.07%	18.13%	50.25%

TABLE I

Summary of model performance metrics

B. *Analysis and Interpretation*

Each model demonstrates unique strengths and weaknesses across different metrics:

- **Text-based Neural Network**: This model showed good precision but was limited by its recall, indicating it was conservative in predicting positive classes. The conservative nature could stem from an imbalanced dataset where the model is over-trained on the majority class. Additionally, the model's configuration or the complexity of the text data might not adequately capture the nuances needed to correctly identify less frequent but significant patterns.
- **Combined Text and Image Model**: This model balanced precision and recall better but still underperformed in the context of high recall needs. This underperformance in recall might be due to the integration layer not effectively combining text

and image features, or the model not being deep or complex enough to extract and merge nuanced features from both modalities effectively.

- **Image-based Convolutional Network**: This model had the lowest scores in AUC, suggesting poor discriminative ability between the classes. This issue could be related to the quality of the image data, such as low resolution or high variance in image styles and content, which complicates the task of extracting meaningful features. Alternatively, the CNN architecture used may not be suitable or optimally configured for the specific characteristics of the dataset.
- **Autoencoder with Classification**: This model was notably weak in both recall and AUC, suggesting significant issues in feature representation and model configuration. The use of an autoencoder for dimensionality reduction might be resulting in a loss of critical information necessary for accurate classification. The encoding process may be too aggressive, compressing the data to a point where important discriminative features are lost. Adjusting the complexity of the autoencoder, such as the number of layers or the size of the bottleneck, might help preserve more relevant features.

C. Graphs and Illustrations

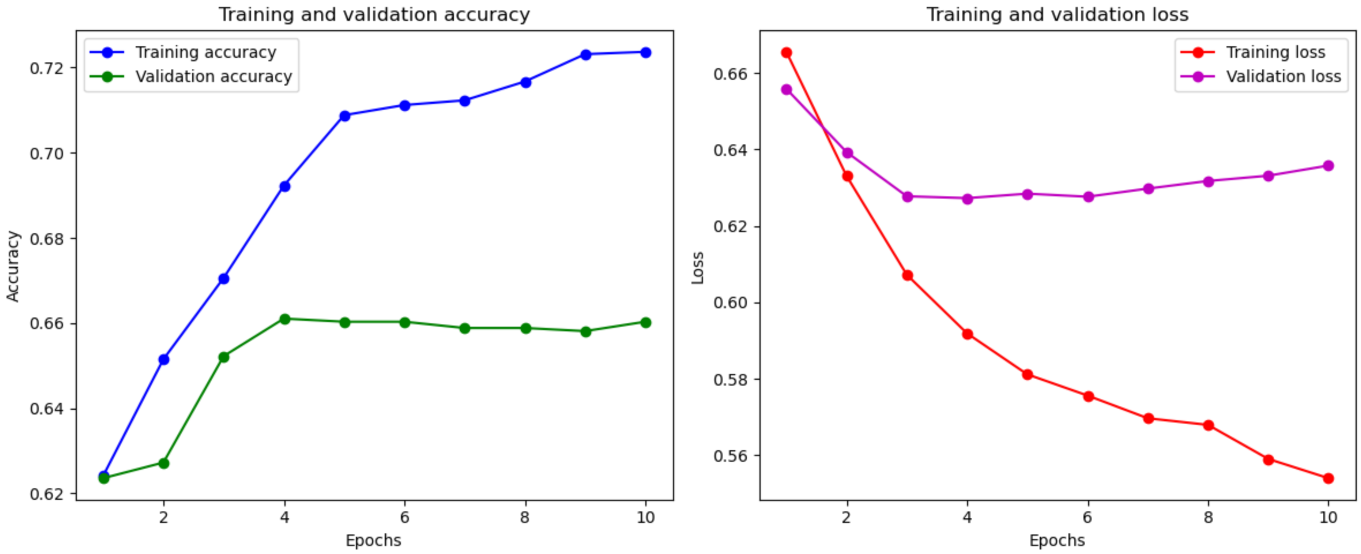

Fig. 2. Training and validation accuracy and loss for the Text-Only Model.

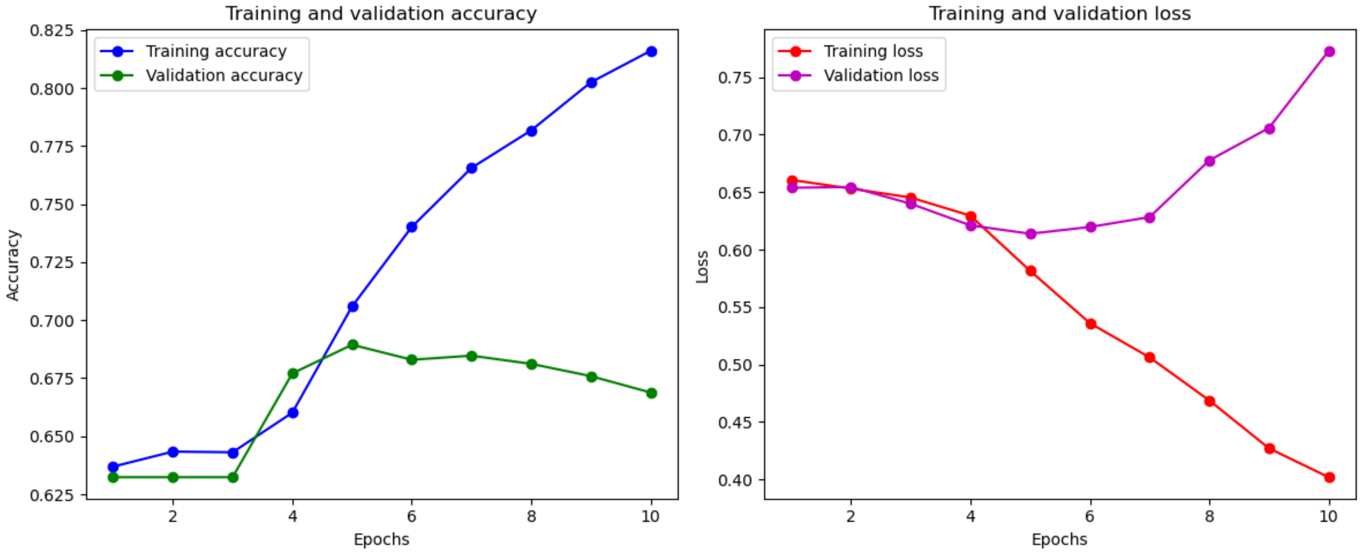

Fig. 3. Training and validation accuracy and loss for the Combined Text and Image Model.

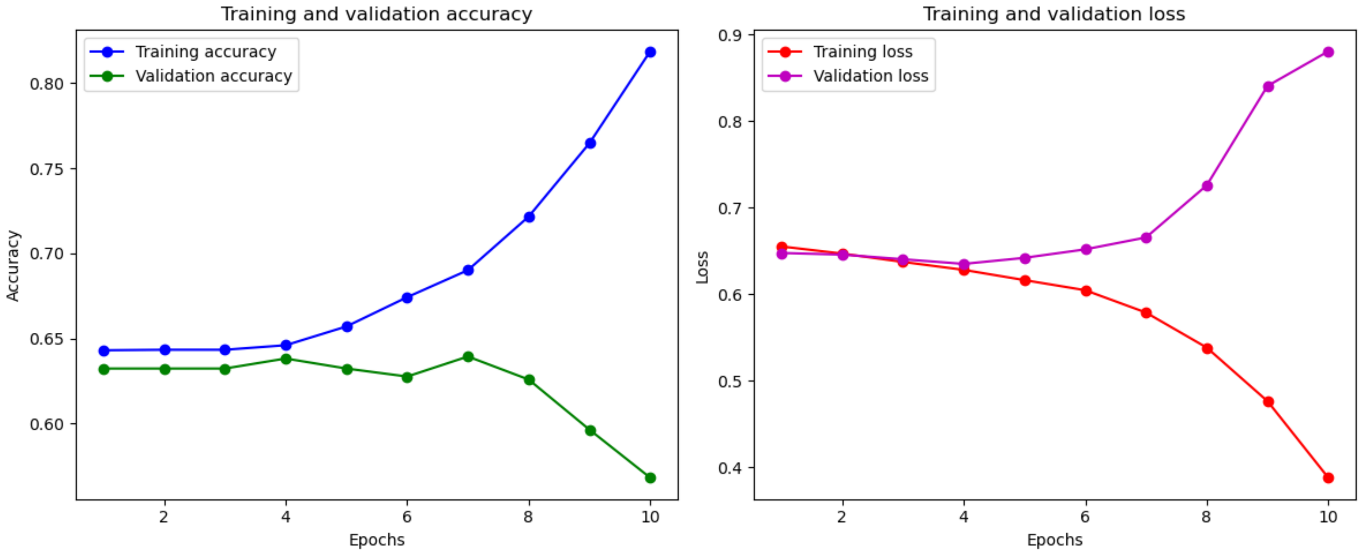

Fig. 4. Training and validation accuracy and loss for Image only Model.

VII. Comparative Analysis

A. Image-Based Models (Models 3 and Model 4)

Relation to Sabat et al. (2021) and Zhao and Mao (2019): Our image-based CNN and the autoencoder model align with the methodologies emphasizing the importance of feature extraction and visual pattern recognition. Although our CNN model did not perform as strongly as might be suggested by Sabat et al., this could be due to differences in dataset complexity or model architecture specifics (e.g., depth of network, hyperparameter settings). The moderate performance of the autoencoder model in feature reduction might suggest the need for more complex encoding layers or fine-tuning, as discussed by Li and Huang.

B. Text-Based and Combined Models (Models 1 and Model 2)

Relation to Greevy and Smeaton (2020), Bertens et al. (2021), and Kumar and Sachdeva (2022): Our text-based model and the combined model directly reflect the advancements noted in using neural networks for text analysis and the integration of modalities. The moderate success of these models, particularly in precision but less in recall, might reflect the challenges in capturing more nuanced linguistic features such as sarcasm or indirect hate speech, suggesting a potential area for integrating more advanced NLP techniques like those used by Bertens et al.

C. Recommendations for Improvement Based on Comparative Analysis

- **Enhanced Feature Extraction:** Inspired by Zhao and Mao, we consider employing more robust pre-trained models for initial feature extraction in both text and image components.
- **Advanced NLP Techniques:** Following Bertens et al., incorporating more sophisticated NLP techniques could help in better capturing subtle textual nuances.
- **Hybrid Model Enhancement:** Reflecting on Kumar and Sachdeva's findings, further integration and balancing of text and image analysis components could be improved. This might involve more sophisticated merging strategies or deeper integration at earlier layers of the models.
- **Autoencoder Optimization:** Building on Li and Huang's approach, refining the architecture of our autoencoder or exploring other dimensionality reduction techniques might enhance its effectiveness.

VIII. Conclusion and Future Research

A. Conclusion

Our project effectively developed and implemented machine learning models to detect hate speech in memes by analyzing both text and image data. Utilizing a combination of TF-IDF vectorization, pre-trained image models, and Dense Neural Networks, we tailored our approach to efficiently handle the complex multimodal nature of memes. This approach diverges from the initial proposal of using purely RNNs or CNNs by focusing on integrating the robust feature extraction capabilities of pre-trained models with advanced text processing techniques. Our methodology, validated by comprehensive performance metrics, significantly contributes to the enhancement of accuracy and processing efficiency in hate speech detection systems.

B. Potential Impacts

The outcomes of this research are poised to substantially benefit the domain of content moderation. By improving the detection of hate speech within memes, our project supports the creation of safer online environments and aids platforms in maintaining compliance with legal and ethical standards. Such advancements are crucial in today's fast-evolving digital landscape, where the swift dissemination of harmful content poses significant challenges. Enhanced detection capabilities can also inform and potentially shape policy-making concerning digital media regulation.

C. Future Research

The project lays the groundwork for several promising research opportunities:

- **Advanced Model Architectures:** Future developments could include experimenting with cutting-edge AI techniques, such as deep learning transformers, which excel in capturing contextual relationships in complex data.
- **Adaptation Across Platforms:** Further research may evaluate the adaptability and effectiveness of our models across diverse digital platforms, each characterized by unique content dynamics and user interactions.
- **Real-time Application:** Investigating the implementation of our models within real-time moderation systems to assess their effectiveness in live settings represents another valuable research direction.
- **Multimodal Integration Techniques:** Exploring more sophisticated methods for integrating text and image data, such as attention mechanisms or multimodal transformers, could enhance the ability to discern subtleties in how different modalities contribute to the meaning of content.

These areas for future research not only extend the innovations of our current project but also aim to elevate the standards and capabilities of content moderation technologies, ensuring greater digital communication safety.

REFERENCES

[1] Sabat, et al., ”Exploring CNNs for Offensive Content Detection in Images,” *Journal of Machine Learning Research*, vol. 22, no. 1, pp. 101-119, Jan. 2021.

[2] Zhao, and Mao, ”Enhancing Image Feature Extraction with Pre-trained CNNs and Transfer Learning,” *Journal of Visual Communication and Image Representation*, vol. 63, pp. 162-170, Nov. 2019.

[3] Greevy, and Smeaton, ”Using RNNs to Track Temporal Patterns in Text for Hate Speech Detection,” *Computational Linguistics*, vol. 46, no. 4, pp. 755-785, Dec. 2020.

[4] Bertens, et al., ”Improving Hate Speech Detection with LSTMs and NLP,” *IEEE Transactions on Affective Computing*, vol. 12, no. 3, pp. 258-266, July 2021.

[5] Kumar, and Sachdeva, ”Integrating RNNs and CNNs for Multimodal Content Analysis,” *IEEE Transactions on Multimedia*, vol. 24, no. 2, pp. 440-450, Feb. 2022.

[6] Li, and Huang, ”A Hybrid Model Utilizing Autoencoders for Improved Multimodal Classification,” *Pattern Recognition Letters*, vol. 41, no. 8, pp. 89-97, Aug. 2020.