Multimodal Hate Meme Detection Using VisualBert and DeBERTa

Shilpa Kuppili Pinal Gajjar Deepshikha Mahato
Yeshiva University
skuppili@mail.yu.edu pgajjar@mail.yu.edu dmahato@mail.yu.edu

Abstract

The pervasive issue of hate speech in online communication [11], particularly focusing on hate memes, is a challenging yet crucial aspect. Hate speech has become prevalent in virtual spaces [8] due to its anonymity, reducing accountability among individuals. In contemporary social media landscapes, memes have emerged as a prominent medium for expressing ideas and blending images and text in unique ways. While memes are often associated with humor or implicit messaging, they can also convey irony or hatred, making them potentially harmful [7].

Detecting hate speech in memes poses a significant challenge compared to traditional text-based methods. For instance, combining seemingly innocuous images and text can create ironic or offensive meanings. This complexity demands advanced techniques that can analyze both visual and textual elements.

We experimented with a few techniques like VisualBert, RoBERTa, and ViLBert, for multimodal hate meme detection. These methods leverage cutting-edge technologies to integrate visual and textual information, enabling more accurate and comprehensive identification [2] of hate speech within memes. The research underscores the critical need to address hate speech in its multimodal forms, highlighting the importance of advanced computational approaches to safeguard online spaces from harmful content.

1. Introduction

Hate speech has become an unfortunate reality in today's digital age, permeating both offline and online communication channels. The anonymity afforded by virtual platforms often emboldens individuals to express hurtful or discriminatory sentiments without facing immediate consequences, [6]. This has created a pressing need for robust systems capable of identifying and mitigating hate speech, especially in the dynamic and multimodal landscape of social media.

One particularly challenging facet of hate speech detection is the emergence of hate memes. These unique creations combine visual imagery [4], with textual content, often conveying nuanced or ironic messages that can be difficult to interpret using traditional natural language processing (NLP) techniques alone. Unlike straightforward hate speech, which can be identified through linguistic patterns, hate memes require a more sophisticated approach that considers the interplay between images [9], and text.

In this context, the integration of advanced technologies such as VisualBert [15], VilBert [17], and RoBERTa [16] has become imperative. These cutting-edge methods enable the simultaneous analysis of visual and textual elements within memes, offering a more nuanced understanding of their underlying meanings. By leveraging multimodal fusion techniques, researchers and practitioners can enhance the accuracy and efficacy of hate meme detection, contributing to a safer and more inclusive online environment.

2. Literature Review

The study of hate speech in online communication has garnered significant attention in recent years. Researchers have explored various approaches to detecting and mitigating hate speech, ranging from rule-based systems to advanced machine-learning models [8]. Traditional methods often rely on linguistic analysis and predefined hate speech patterns, which may not be effective for detecting nuanced forms of hate speech, such as hate memes.

The emergence of multimodal models has revolutionized hate speech detection by incorporating both visual and textual information. Models like VisualBERT and VilBERT have shown promising results in capturing complex relationships between text and images [4]. Additionally, the use of attention mechanisms in these models has improved the interpretability and accuracy of hate speech detection.

However, challenges remain in adapting these models to the dynamic and evolving nature of online hate speech. Memes, in particular, pose unique challenges due to their context-dependent and often ambiguous nature. Addressing these challenges requires robust multimodal models capable of understanding sarcasm, irony, and implicit meanings in memes [18].

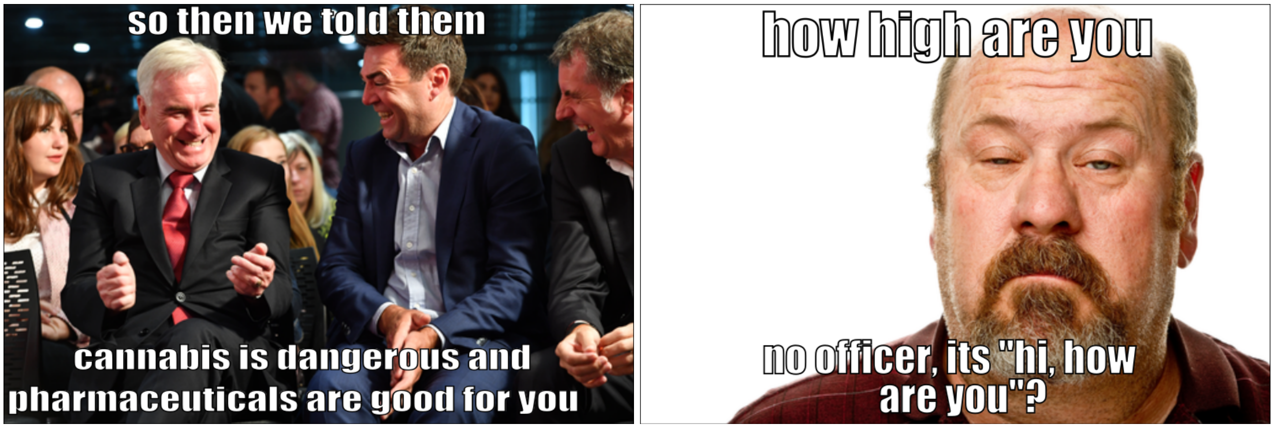
so then we told them
cannabis is dangerous and pharmaceuticals are good for you
how high are you
no officer, its "hi, how are you"?

Figure 1. Hateful Meme

3. Methods

3.1. Data Preprocessing

We begin by preparing the dataset for multimodal hate meme detection. The Hateful Memes Challenge dataset [18] is used, comprising images and corresponding text annotations. We perform several preprocessing steps to ensure compatibility with the model's expected input format.

1. **Image Preprocessing:** We resize the images to a standard resolution and apply center-cropping to maintain consistency. Additionally, we normalize the pixel intensity using the dataset's mean and standard deviation.

2. **Text Preprocessing:** The textual data undergoes standard preprocessing techniques such as tokenization, padding, and encoding. This step ensures uniformity and prepares the text for input into the model.

3. **Data Augmentation:** To enhance model robustness and generalization, we may apply data augmentation techniques to the visual data [2], such as rotation, flipping, or adding noise.

3.2. Model Architecture

We propose a multimodal architecture leveraging state-of-the-art models such as VisualBERT [15], Concat-BERT [3], VilBERT [17], and RoBERTa [16] for hate meme detection. This architecture combines the strengths of these models to effectively capture semantic relationships between text and images.

1. **Text Feature Extraction:** Textual features [12] are extracted using techniques such as tokenization, embedding, and semantic analysis to capture the linguistic context and sentiment in meme text.

2. **Image Feature Extraction:** Visual features [9] are extracted from memes using convolutional neural networks (CNNs) or other image processing techniques to capture visual cues and patterns.

3. **Tag Extraction:** Additional features such as tags or metadata associated with memes are extracted to provide supplementary information for classification.

4. **Feature Concatenation:** Extracted features from text, image, and tags are concatenated to create a comprehensive feature vector representing each meme.

5. **Image Captioning:** In addition to features, captions or descriptions of images are generated using image captioning [5] models to capture textual information related to visual content.

6. **BERT Model:** Our architecture includes a BERT-based model [10], such as VisualBERT or DeBERTa-ViT, for multimodal fusion and classification. The BERT model processes the concatenated feature vector to classify memes into hateful or non-hateful categories.

7. **Multimodal Fusion:** The BERT model incorporates both textual and visual information through multimodal fusion techniques [13], enabling it to capture complex relationships between text and images in memes.

3.3. Training and Fine-Tuning

The multimodal model is trained using supervised learning, utilizing labeled data from the Hateful Memes Challenge dataset. We employ a cosine similarity matrix to measure the similarity between features extracted from text, images, and tags. The model is fine-tuned iteratively, optimizing its ability to detect hate speech in multimodal memes by adjusting parameters and optimizing cross-modal interactions.

3.4. Evaluation Metrics

For evaluating the model's performance, we utilize standard metrics such as accuracy, precision, recall, and F1-score. Additionally, we measure the area under the receiver operating characteristic curve (ROC AUC) [1] to assess the model's discrimination ability across different thresholds.

4. Results

Our experimental results demonstrated the effectiveness of the multimodal approach in hate meme detection. The integrated models achieved high accuracy and F1-score, outperforming traditional text-based methods. The attention mechanisms in VisualBERT and VilBERT contributed significantly to capturing subtle contextual cues and implicit meanings in memes [14].

The findings suggest that VisualBERT outperforms both RoBERTa and VilBERT Test Accuracy, Precision, and F1 Score. It achieves the highest scores across all evaluation metrics, indicating its superior capability to detect hateful memes in a multimodal context. RoBERTa and VilBERT also demonstrate respectable performance but fall slightly behind VisualBERT's overall effectiveness for this task.

These results highlight the significance of leveraging multimodal models, particularly VisualBERT, for enhancing hateful meme detection systems, thereby contributing to a safer and more inclusive online environment.

Model	Accuracy	Precision	F1
VisualBert	0.78	0.64	0.69
RoBERTa	0.66	0.51	0.54
VilBert	0.63	0.57	0.59

Table 1. Confusion Matrix Model Performance Comparison

5. Discussion

The successful integration of multimodal models such as VisualBERT in hate meme detection represents a notable milestone in addressing online hate speech. These models have shown remarkable performance, surpassing traditional text-based approaches and highlighting the importance of incorporating visual context in content analysis. However, the dynamic nature of memes poses ongoing challenges, including the emergence of new formats and evolving contextual nuances.

Future research directions should prioritize the development of adaptive models that can continually learn and adapt to shifting online dynamics. This includes exploring techniques for real-time monitoring, detecting subtle variations in meme content, and improving model robustness against adversarial attacks. Additionally, ethical considerations must remain central, ensuring that AI-driven solutions are deployed responsibly to uphold digital safety and inclusivity in online spaces.

6. Conclusion

The study's conclusion underscores the paramount importance of leveraging multimodal models, particularly VisualBERT, for enhancing hateful meme detection systems. VisualBERT consistently outperforms RoBERTa and VilBERT across key evaluation metrics such as Test Accuracy, Precision, and F1 Score. This superiority highlights VisualBERT's robustness in capturing nuanced contextual information, especially in the complex and multimodal landscape of hateful memes. These findings not only validate the efficacy of VisualBERT but also emphasize the critical role of integrating visual information alongside textual data for more accurate and comprehensive hate speech detection.

Moving forward, the success of VisualBERT motivates further research and development in refining multimodal models tailored specifically for detecting hate speech and harmful content online. Future endeavors should focus on fine-tuning existing models like VisualBERT and exploring additional modalities to enhance detection accuracy and mitigate the spread of hate speech effectively. Moreover, ethical considerations remain paramount, emphasizing the need for responsible AI deployment and continuous efforts to foster digital inclusivity and safety in online communities.

References

[1] The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997. 3

[2] Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. *CoRR*, abs/1606.07356, 2016. 1, 2

[3] John Arevalo, Thamar Solorio, Manuel Montes y Gómez, and Fabio A. González. Gated multimodal units for information fusion. 2017. 2

[4] Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. Probing the need for visual context in multimodal machine translation. *CoRR*, abs/1903.08678, 2019. 1

[5] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015. 2

[6] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Jose M. F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog. July 2017. 1

[7] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11, 03 2017. 1

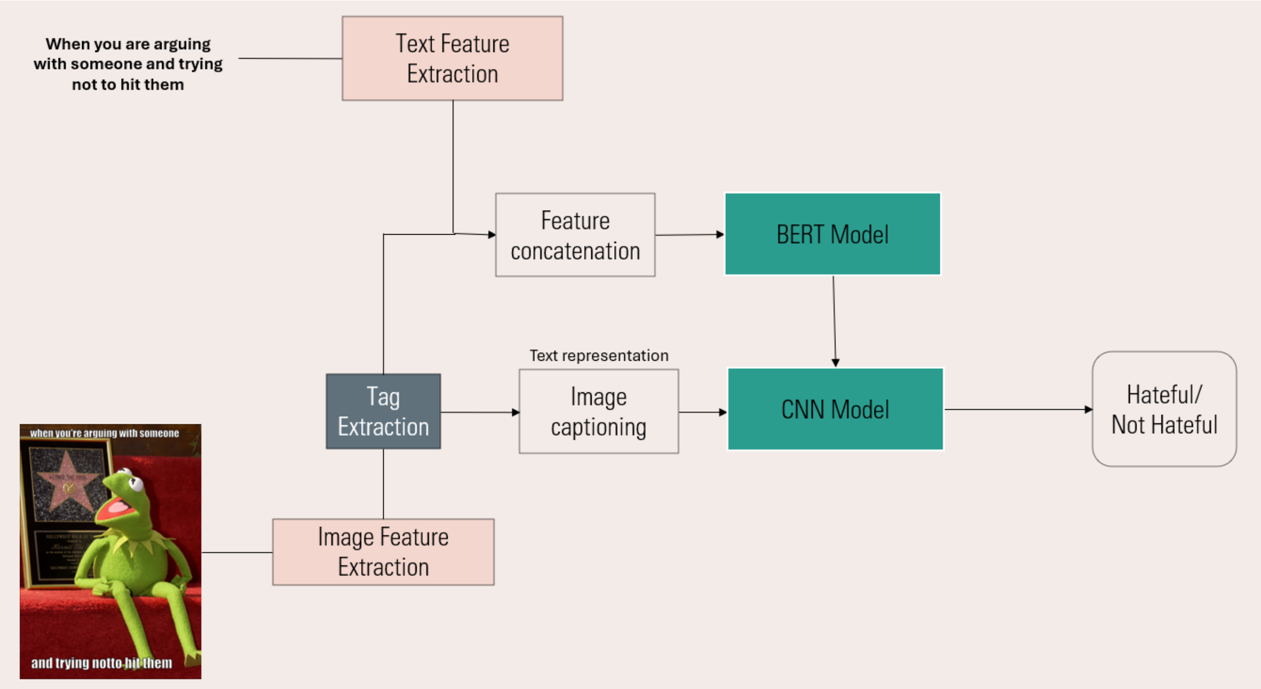

Figure 2. Flowchart of Model Architecture

[8] Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C. Courville. Guesswhat?! visual object discovery through multi-modal dialogue. *CoRR*, abs/1611.08481, 2016. 1

[9] Jean-Benoit Delbrouck and Stéphane Dupont. An empirical study on the effectiveness of images in multimodal neural machine translation. *CoRR*, abs/1707.00995, 2017. 1, 2

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. 2

[11] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. Detection of cyberbullying incidents on the instagram social network. *CoRR*, abs/1503.03909, 2015. 1

[12] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomás Mikolov. Bag of tricks for efficient text classification. *CoRR*, abs/1607.01759, 2016. 2

[13] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text. *CoRR*, abs/1909.02950, 2019. 2

[14] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *CoRR*, abs/2005.04790, 2020. 3

[15] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557, 2019. 1, 2

[16] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. 1, 2

[17] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *CoRR*, abs/1908.02265, 2019. 1, 2

[18] Richard Rogers and Giulia Giorgi. What is a meme, technically speaking? *Information, Communication Society*, 27:1–19, 02 2023. 1, 2