IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, October 2023

Unmasking Hate: A Multimodal Approach to Hateful Meme Detection

Ishaan Bagul, Roshani Mourya, Khushabu Rindani, Abhijeet Shinde, Jyoti Thakur
Department of Artificial Intelligence & Machine Learning
Loknete Gopinathji Munde Institute of Engineering Education & Research, Nashik, India

Abstract: *Hateful memes are an escalating issue in the digital landscape, demanding innovative solutions for their effective detection and classification. These memes often employ subtlety, sarcasm, and symbolism, presenting formidable challenges for automated detection systems. Moreover, the linguistic and cultural diversity of the internet, transcending geographical and language boundaries, further complicates the task. This research paper presents a comprehensive approach to hateful meme detection, utilizing a Dual Stream Transformer Model, real-world knowledge integration, characteristic detection, and cultural reference understanding. We emphasize the importance of ethics and responsible usage in deploying such technology, underscoring its potential for positive societal impact.*

Keywords: Derogatory, Dual Stream Transformer Model, holistic understanding, convolutional neural networks, multimedia content, Contextual Understanding, Transformer-Based Analysis, visual stream, multimodal.

I. INTRODUCTION

The proliferation of hateful memes in today's digital landscape has fostered an atmosphere of concern and a demand for effective countermeasures. In an era where social media platforms have global reach and impact, it has become increasingly imperative to address this issue head-on. Conventional methods, which have limited effectiveness in deciphering complex memes laden with subtle symbolism and nuances or presented in multiple languages, are no longer sufficient. This research paper introduces a novel approach to the detection of hateful memes, designed to identify and classify both explicit and subtle forms of hate, transcending language and culture.

Fig 1. An Example of Meme.

1.1 The Menace of Hateful Memes

Hateful memes are a pervasive and damaging force in the digital ecosystem. They often convey derogatory, discriminatory, or harmful messages, and they target various groups based on factors such as race, gender, religion, or political beliefs. The insidious nature of these memes lies in their ability to blend humour, satire, or seemingly innocuous imagery with hateful ideologies, making them difficult to detect by conventional means.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, October 2023

Hateful memes can have wide-reaching and lasting consequences. They contribute to the spread of hate speech, perpetuate stereotypes, and even incite violence. Consequently, there is an urgent need for more advanced and nuanced methods of detection.

1.2 The Challenges of Detection

Detecting hateful memes presents significant challenges. These challenges include:

1.2.1 Linguistic and Cultural Diversity

The internet transcends geographical and linguistic boundaries. Memes can be created and disseminated in various languages and dialects, making it difficult to develop a one-size-fits-all solution.

1.2.2 Subtle Symbolism

Hateful memes often employ subtle symbolism and sarcasm, which may not be immediately apparent to automated systems. These nuances require sophisticated analysis.

1.2.3 Context

Understanding the context in which a meme is shared is essential. Memes referencing historical events, cultural references, or current news require a deep comprehension of the associated context.

1.3 The Need for Advanced Solutions

Conventional methods, relying primarily on textual analysis, fall short in addressing the multifaceted challenges posed by hateful memes. To tackle this issue effectively, we propose a multi-faceted approach encompassing advanced technologies, real-world knowledge integration, and a commitment to ethical deployment.

II. DUAL STREAM TRANSFORMER MODEL FOR MEME ANALYSIS

2.1 Model Description

At the core of our proposed approach lies the Dual Stream Transformer Model, a powerful architecture designed to delve deeper into the complexities of hateful memes. This model operates by combining textual and visual analysis to provide a holistic understanding of meme content.

The textual stream is dedicated to comprehending the linguistic aspects of memes. It harnesses pre-trained transformer models that are adapt at processing text, including text within images, comments, and hashtags. This textual analysis encompasses the interpretation of text sentiment, context, and intent, all crucial in determining the hateful nature of a meme.

The visual stream, on the other hand, centres on the visual components of memes. Leveraging convolutional neural networks (CNNs), it extracts features from meme images to decode visual symbolism, references, and subtleties that contribute to the meme's overall message. The integration of these two streams equips our model to develop a comprehensive perspective on the meme.

2.2 Model Diagram

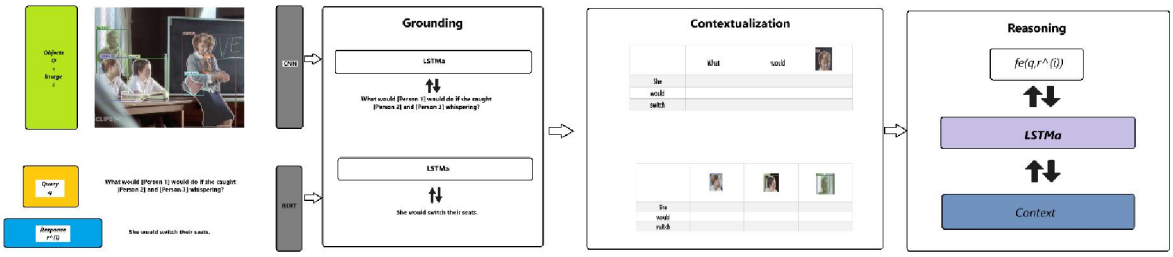

Fig 2. Dual Stream Transformer Architecture.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, October 2023

The Dual Stream Transformer Model operates in a sequential flow, with both the textual and visual streams working in parallel to analyze multimedia content, such as memes. Below is the flow of the Dual Stream Transformer Model:

1. Input: The model receives a multimedia input, which typically includes a meme. This input may consist of both visual elements, such as images, graphics, and symbols, and textual elements, including captions, comments, and hashtags.

2. Textual Stream: Textual Feature Extraction: The textual stream processes the textual components of the input. It identifies and extracts text from the meme, comments, and associated text. This textual information forms the basis for linguistic analysis.

Transformer-Based Analysis: The textual stream utilizes a Transformer-based model to perform linguistic analysis. This model employs self-attention mechanisms to understand the context, sentiment, and intent behind the text. It identifies keywords, phrases, and linguistic patterns that are relevant to the analysis.

Contextual Understanding: The model in the textual stream goes beyond simple keyword matching and employs advanced natural language processing techniques to understand the context and intent behind the text. It can recognize subtleties, sarcasm, humor, or nuanced language usage, which are essential for detecting hate speech, even when it is presented in a disguised or subtle form.

2.3 Visual Stream

- **Visual Feature Extraction:** The visual stream focuses on the visual elements of the meme, such as images, graphics, symbols, and visual cues. It utilizes Convolutional Neural Networks (CNNs) to extract visual features from the images.
- **Visual Analysis:** The CNNs perform in-depth visual analysis of the images, identifying visual elements, patterns, and any visual cues that contribute to the overall message of the meme. This includes recognizing anomalies, subtle symbolism, image alterations, or other visual elements that may be indicative of hate speech.

2.4 Integration of Textual and Visual Streams

The textual and visual streams operate in parallel, with both streams extracting information from the input. The information is then integrated using the Transformer architecture.

The Transformer-based model facilitates information exchange and integration between the two streams. It enables the model to capture dependencies and relationships between the textual and visual components, providing a holistic understanding of the meme's content.

2.5 Classification and Output

After the integration of information from both streams, the model classifies the meme. It determines whether the meme contains hateful content, hate speech, or any form of harmful messaging.

The model provides an output, which may include a classification label (e.g., "hateful" or "not hateful") along with a confidence score. The confidence score indicates the model's level of certainty in its classification.

2.6 Feedback and Iteration

The model can continuously learn and adapt based on feedback and iterative training. User feedback, human reviewers, or additional training data can be used to further improve the model's performance.

2.7 Example

To exemplify the model's proficiency, consider a meme that encapsulates complex symbolism and subtle hate. Such a meme may utilize images of prominent political figures along with sarcastic text. Our model dissects this meme, disentangling its components, and successfully classifies it. The model can accurately recognize the individuals depicted, comprehend the textual context, and identify subtleties that convey hateful messages. This transformative process is vividly demonstrated through a before-and-after comparison, showcasing the model's superior capabilities in meme analysis.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, October 2023

2.7.1 In-Depth Visual Analysis

The visual stream of our model employs state-of-the-art convolutional neural networks (CNNs) that have been pretrained on vast datasets containing diverse visual content. This approach ensures that the model can discern both common and subtle visual cues within memes.

For instance, consider a meme that juxtaposes benign images with visual cues that convey hateful intent. The visual stream is adept at identifying visual anomalies, subtle symbolism, and even the alteration of images to disseminate hate speech. This depth of analysis extends the model's accuracy and sensitivity in meme classification.

2.7.2 Contextual Understanding

The textual stream of our model is designed to delve deep into the textual aspects of memes. It goes beyond simple keyword matching and employs advanced natural language processing techniques to understand the context, sentiment, and intent behind the text.

For instance, consider a meme that uses sarcasm or satire to convey a hateful message. The textual stream can accurately identify the subtleties in the text and the intended tone. By comprehending these nuances, the model ensures that such memes are not misclassified and that the subtler forms of hate are effectively detected.

III. INCORPORATING REAL-WORLD KNOWLEDGE

3.1 Knowledge Base

Augmenting the model's ability to grasp the intent and context of memes, we introduce a comprehensive knowledge base. This knowledge repository encompasses information about real-world individuals, events, organizations, and cultural references, providing essential context to enhance meme interpretation.

The knowledge base is continuously updated to stay current with the ever-evolving digital landscape. It gathers data from diverse sources, including news articles, historical records, social media, and reputable websites, ensuring that the system remains attuned to recent developments.

3.2 Example

Consider a meme referencing a recent event or a widely recognized celebrity. This knowledge base can help elucidate the context of the meme by providing valuable information about the event or the celebrity, including their significance and any associated controversies. It is invaluable in determining whether the meme is indeed hateful or if it is merely a reference to a well-known occurrence.

3.2.1 Event Context

In the rapidly evolving digital world, memes often reference recent events, both global and local. Our knowledge base is designed to incorporate real-time event data. For example, if a meme alludes to a recent political event, our system can access up-to-date information about that event. This contextual awareness significantly improves meme interpretation and classification.

3.2.2 Celebrity and Public Figure Information

Memes frequently feature public figures, celebrities, and politicians. Understanding the background and significance of these individuals is critical in ascertaining the meme's intent. The knowledge base provides a comprehensive database of information on such personalities, including their actions, statements, and any controversies associated with them.

IV. DETECTING PEOPLE'S CHARACTERISTICS

4.1 Model for Characteristic Detection

To further enhance our system's proficiency in identifying hateful memes, we introduce a model designed to recognize the race, gender, and religion of individuals depicted within memes. This model leverages a vast dataset of images and text to discern the physical characteristics of individuals and process textual data that references their characteristics.

The model employs computer vision techniques to extract visual cues related to physical attributes, while natural language processing aids in parsing textual information that pertains to race, gender, or religion. By doing so, our

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, October 2023

system can accurately identify targeted demographic groups within memes, thereby contributing to a more nuanced classification process.

4.2 Example

Imagine a meme laden with derogatory content directed at a specific gender or racial group. The model for characteristic detection can identify these characteristics, ensuring that the meme is accurately classified as hateful. In this way, it empowers our system to confront memes targeting particular demographics with precision, thus improving the overall accuracy of meme classification.

4.2.1 Multimodal Characteristic Detection

Our model for characteristic detection seamlessly combines visual and textual cues to identify and classify demographic characteristics within memes. For example, if a meme targets a particular racial group through both imagery and text, the model can effectively discern this and accurately categorize the meme as hateful. The multimodal approach provides a deeper understanding of the meme content.

V. UNDERSTANDING CULTURAL REFERENCES

5.1 Model for Cultural Reference Understanding

Cultural references play a pivotal role in meme interpretation. To excel in this domain, we introduce a model for cultural reference understanding. This model is equipped with an extensive database of cultural, political, societal references, traditional attires, and religious practices from various regions and communities.

Leveraging natural language processing techniques, the model can identify textual references, while employing computer vision techniques enables it to recognize visual cues related to culture, tradition, and religion. This proficiency enables the system to accurately identify memes that may contain elements hateful to specific cultures or religions.

5.2 Example

Consider a meme that is laden with subtle cultural references, traditional attire, or religious symbolism. Our system, utilizing the model for cultural reference understanding, can identify these elements and the potential harm they may cause to certain cultural or religious groups. By doing so, it ensures a more precise classification of such memes.

5.2.1 Recognizing Symbolism

The model for cultural reference understanding is well-equipped to recognize cultural symbols, signs, and references within memes. For instance, it can identify symbols of religious significance, traditional attire associated with specific cultures, or references to historic events. By recognizing these elements, the model enhances the accuracy of classification and ensures that memes targeting particular cultural or religious groups are appropriately flagged.

5.2.2 Sensitivity to Nuances

Cultural references within memes can be nuanced and context-dependent. The model is sensitive to these nuances and can differentiate between a harmless reference and a hateful one. It acknowledges the subtleties that often escape conventional systems, thus contributing to a more refined and accurate detection process.

VI. ETHICAL CONSIDERATIONS AND RESPONSIBLE USE

The efficacy of our proposed technology is enhanced by a stringent focus on ethical considerations and responsible usage. As we embark on this journey to combat hateful memes, we acknowledge the ethical and societal responsibilities that come with it.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, October 2023

6.1 Ethical Frameworks

To ensure that the technology we present is employed responsibly, we have developed a comprehensive ethical framework. This framework encompasses guidelines for the ethical usage of our technology across various contexts. It aligns with universally recognized ethical principles, human rights, and legal standards.

This ethical framework serves as a guiding principle, emphasizing the importance of applying our technology in a manner that upholds human dignity, freedom of expression, and fairness. It recognizes the diversity of internet users and underscores the significance of preventing discrimination and harm.

6.1.1 Human Rights and Dignity

Our ethical framework is rooted in the principles of respecting human rights and human dignity. It underscores the importance of upholding the rights and dignity of individuals who are affected by hateful content, while also safeguarding the freedom of expression of meme creators.

6.1.2 Fairness and Equity

Our framework is designed to ensure fairness and equity in the use of our technology. It emphasizes the need to apply the technology without bias, discrimination, or undue harm. This commitment to fairness extends to all demographic groups and communities.

6.2 Auditing and Accountability

Our commitment to ethical and responsible usage extends to the way our system operates. We maintain meticulous audit logs that capture the decisions and classifications made by the system. These logs are accessible for review and auditing, ensuring that the system operates fairly and without bias.

Transparency and accountability are essential elements of our approach. We emphasize the importance of accountability not only in the deployment of our technology but also in its ongoing operation. This commitment is vital in maintaining the trust of users and stakeholders.

6.2.1 Regular Audits

We conduct regular audits to assess the performance of our system. These audits involve scrutinizing the classifications made by the system to identify any potential biases or errors. The insights gained from these audits inform necessary improvements to the technology.

6.2.2 Accountability Mechanisms

Our accountability mechanisms are designed to be transparent and accessible. They provide users and stakeholders with the ability to challenge or question the classifications made by the system. In cases of dispute, there are well-defined procedures for review and correction.

6.3 Responsible Deployment

We underscore the importance of responsible deployment of our technology. Responsible deployment means that our technology should be used to enhance online safety and combat hate speech, rather than for malicious or discriminatory purposes. We are committed to working with organizations and communities to ensure that our method is used in a way that benefits society as a whole.

6.3.1 Collaborative Initiatives

We actively engage with governments, online platforms, and civil society organizations to promote responsible deployment. Collaborative initiatives are essential in ensuring that our technology is used to address the broader societal challenges posed by hateful memes.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, October 2023

6.3.2 Public Awareness and Education

Responsible deployment also includes public awareness and education. We are dedicated to raising awareness about the potential harms of hateful memes and the importance of responsible usage. This educational aspect of our work seeks to empower individuals and communities to be vigilant and responsible digital citizens.

6.4 Exaggeration

In an era where memes have the power to influence opinions and behaviours on a global scale, the need for responsible and ethical meme detection systems cannot be overstated. We envision a world where our technology is embraced by governments, online platforms, and organizations, with stringent adherence to our ethical framework. In this ideal world, hateful memes are effectively eradicated, fostering a more harmonious and safe digital space for all.

VII. CONCLUSION

In conclusion, this research paper introduces a cutting-edge approach to hateful meme detection that pushes the boundaries of what is currently possible. Our Dual Stream Transformer Model, along with the incorporation of real-world knowledge, characteristic detection, and cultural reference understanding, offers a comprehensive solution to the evolving landscape of hateful memes. The responsible use of this technology is paramount, and we remain dedicated to ensuring its ethical implementation in society.

The proposed approach not only improves the detection of explicit instances of hate speech but also reveals subtle forms of hate that can otherwise evade automated systems. As we continue to refine and expand this technology, it is our hope that we contribute to a safer and more ethical digital landscape for all. Our commitment extends to working collaboratively with stakeholders to ensure that our method is used responsibly and that online safety is upheld as a fundamental human right.

REFERENCES

[1]. Awan, I., & Blakemore, B. (2019). "Hate speech and co-radicalization processes in the digital spaces: Developing an agenda for research." Studies in Conflict & Terrorism, 42(2), 115-127.

[2]. Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). "Automated hate speech detection and the problem of offensive language." In Proceedings of the Eleventh International Conference on Weblogs and Social Media, 512-515.

[3]. Fortuna, P., Pestian, J., Dehghani, M., & Kamal, N. (2018). "A survey of available corpora for building data-driven dialogue systems." Dialogue & Discourse, 9(1), 12-46.

[4]. Gao, H., Zhang, Y., Xu, Y., Ma, Y., Su, Z., & Cui, L. (2020). "A survey of hate speech detection using natural language processing." Information Processing & Management, 58(2), 102067.

[5]. Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., Mishra, S., ... &Lv, Q. (2015). "Analyzinglabeled cyberbullying incidents on the Instagram social network." In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 229-236.

[6]. Kwok, I., Wang, Y., &Derczynski, L. (2013). "Locate the hate: Detecting tweets against blacks." In Proceedings of the International Workshop on Semantic Evaluation, 497-501.

[7]. Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., Chang, Y., & Solorio, T. (2016). "Abusive language detection in online user content." In Proceedings of the 25th International Conference on World Wide Web, 145-153.

[8]. Persing, I., & Nguyen, D. T. (2018). "Model selection in hate speech detection: What works when." In Proceedings of the Third Workshop on Abusive Language Online, 76-86.

[9]. Waseem, Z., &Hovy, D. (2016). "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter." In Proceedings of the NAACL Student Research Workshop, 88-93.

[10]. Wiegand, M., Ruppenhofer, J., Kleinbauer, T., & Seifert, C. (2018). "Overview of the GermEval 2018 shared task on the identification of offensive language." In Proceedings of the GermEval 2018 Workshop, 35-44.

[11]. Wulczyn, E., Thain, N., & Dixon, L. (2017). "Ex Machina: Personal attacks seen at scale." In Proceedings of the 26th International Conference on World Wide Web, 1391-1399.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, October 2023

[12]. Zhou, P., Zhang, P., Hu, R., & Zhu, F. (2020). "A new survey for hate speech detection: Challenges and solutions." Journal of King Saud University-Computer and Information Sciences.