Real Memes In-The-Wild: Explainable Classification of Hateful vs. Non-Hateful Memes

Rohan Singh Leekha, Department of Computer Science, Virginia Tech, USA

Eugenia H. Rho, Department of Computer Science, Virginia Tech, USA

1 BACKGROUND

Functioning as visual punch-lines embedded with humor and satire, memes have become a mainstay of online public discourse [18]. Memes function as "units of cultural transmission" [22] and can quickly spread or reinforce ideas across online communities [13,22]. Recent studies [4,18] have shown that people sometimes use memes to disseminate violent [12], hateful [6], and misogynistic messages disguised as humor [6,15]. The virality of such harmful memes over the recent years has encouraged deep learning (DL) research on hateful meme classification. These DL models however, are exclusively trained to classify memes based on synthetically generated data [1,7,21]. Synthetically generated meme data, such as the widely used Hateful Memes Challenge dataset from Meta AI was created by interchanging random texts with random images. Such artificially generated memes often exclude neologisms, insider- expressions, slangs and other linguistic nuances, which are prevalent across *real* memes that actually circulate online [19,22]. As a result, current state-of-the-art classifiers perform poorly when tasked to predict real hateful memes [17]. Furthermore, such studies tend to focus solely on the prediction task rather than explaining the characteristics that make memes hateful– meaning, classification results typically lack any explanation as to why a meme is predicted as hateful [1,2,9,17]. We aim to address these gaps. First, we share a manually curated in-the-wild hateful meme dataset, *RealMemes* (3,142 memes) collected from Instagram, Reddit, and WhatsApp and Telegram groups. Unlike collecting textual data, compiling a meme dataset is particularly challenging as the overlaid text on a visual image makes searching through keywords difficult. Second, we make hateful meme classification results explainable by building an interpretable multimodal classification system that not only classifies hateful vs. non-hateful memes, but also identifies key words and visual descriptors associated with hateful vs. non-hateful memes.

2 ANALYSIS

Data Collection. In this study, we manually collected memes from various social media platforms, such as Reddit, Instagram, Facebook, and public WhatsApp and Telegram channels known to contain political satire or humorous memes (Table 1). We first defined what is hateful vs. non-hateful. A hateful meme contains a direct or indirect attack on people based on categorical characteristics, such as ethnicity, race, nationality, religion, caste, sex, gender, and sexual orientation. We define attack as violent or dehumanizing speech (e.g., comparing people to nonhuman entities, such as chattel), statements of inferiority, and calls for exclusion or segregation. Mocking hate crime was also considered hateful. For non-hateful memes, we used the same categorical characteristics, but did not include direct or indirect attacks. Such non-hateful memes did not attack any individual or group. Two authors manually collected and analyzed hateful memes to ensure consistency in labeling. Only memes that were unanimously voted as hateful were considered for analysis. To preprocess our data, we manually cropped the image component and extracted texts using the Google Vision API [14] for all the memes in our dataset.

Social Media Platform	Hateful Memes	Non-Hateful Memes
Reddit	542	1216
Instagram	456	300
Telegram / WhatsApp	237	391
Combined / Total	1235	1907

Table 1: Breakdown of our real in-the-wild meme dataset by source

Classification. We built a multimodal classifier one language and one vision DL model to extract features from text and images in the form of embeddings. For text embeddings, we used the BERT [5] to capture contextual word representations (step 1A) and used the pre-trained VGG-16 [3,8] Convolutional Neural Net (CNN) architecture to extract visual embeddings

(step 1B). Given that multimodal architectures can capture more than one modality of information, they have several advantages, such as improved accuracy, increased robustness, and enhanced interpretability of results [11,20]. Hence, we combined the two models. We concatenated the textual and visual features (step 2) and added three dense layers with ReLu activation (step 3). We added a classification layer with a sigmoid activation function to predict memes as hateful or non-hateful (step 4).

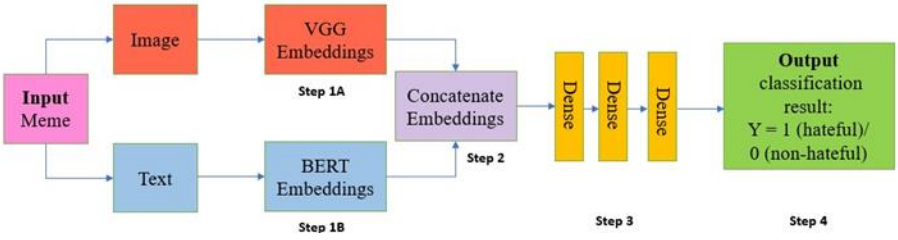

Fig 1. Classification framework to identify hateful and non-hateful memes

Interpretation. We use Integrated Gradients (IG) [16] to identify tokens that are most attributable to hateful vs. non-hateful memes. IG works by calculating the gradient of a model's prediction output to its input features, providing intuitive explanations for output decisions from transformer-based models like BERT [16]. We used the IG attribution scores - ranging from -1(predictive of non-hateful) to 1(predictive of hateful) - to identify tokens that were most predictive of the textual branch of the multimodal classification decision. To identify visual patterns most predictive of hateful memes, we used Gradient SHAP [10]. We calculated how much each individual pixel in the image component of the meme contributed to the classification decision as to whether a meme was hateful vs. non-hateful. The darker the density of the pixels highlighted in an image, the stronger its predictive contribution to the classification decision [10].

3 RESULTS

Our model achieved an accuracy of 90.9%, which was confirmed through 4 cross-fold validation, while bootstrap analysis revealed a 95% confidence interval of 90.1% to 91.2% based on 20% test data (label split: 50/50). **Tokens Predictive of Hateful Memes**. Our findings show that words and phrases most predictive of hateful memes tend to be associated with 1. derogatory references to minorities (*katua*[1], *Bulli*[2], *Mulla*[3] and *jihad*[4]), 2. controversial issues (*Ram*[5], *Kashmir*), religious identity (*Hindu*, *Muslim*, *Islam*), and abusive language (*chutiya*[6]). **Visual Patterns Predictive of Hateful Memes**. Our Grad SHAP interpretation results show that pixels most predictive of hateful memes tend to visually highlight violent imagery. For example, Fig 2a shows a meme of a Nazi soldier character beheading a cartoonized Muslim man (characterized by the skull cap). In the Grad SHAP rendition of this meme shown in Fig 2b, the darkest pixels highlight the act of beheading using a knife, indicating that these pixels are associated with features most predictive of what makes the meme hateful.

(a) (b) (c) (d)

Fig 2. (a) and (c) are two original memes classified as hateful; their corresponding renditions using Grad-SHAP are shown in (b) and (d)

[1] Muslim Man [2]Sexualizing Muslim Women [3]Religious Muslim Man [4]Holy War [5]Hindu God [6]Idiot

Similarly, in Fig 2d, Gradient- SHAP highlights the turban worn by the widely known *Pepe the frog* with darker pixels, indicating that the turban, which is typically associated with Islam or Sikhism is contributive to the model's prediction of this meme as hateful. The identification of the turban in this instance provides insight into the ways in which visual cues can be used to propagate hate against certain minority groups. Through this work, we highlight the importance of addressing both textual and visual content in detecting and mitigating hateful online behavior. The classification of hateful memes based on identifiable visual and textual descriptors can equip content moderators with concrete justification in their decision-making to ensure a more effective and accurate identification and mitigation of hateful online behavior in memes. Identifying such elements also provides opportunities to enhance meme classifiers trained on synthetic data allowing them to better recognize and classify hateful memes in real-world situations. Finally, our "in-the-wild" dataset provides a valuable resource for studying hateful behavior online

REFERENCES

[1] Apeksha Aggarwal, Vibhav Sharma, Anshul Trivedi, Mayank Yadav, Chirag Agrawal, Dilbag Singh, Vipul Mishra, and Hassène Gritli. 2021. Two-way feature extraction using sequential and multimodal approach for hateful meme classification. Complexity 2021, (2021), 1–7.

[2] Piush Aggarwal, Michelle Espranita Liman, Darina Gold, and Torsten Zesch. 2021. VL-BERT+: Detecting Protected Groups in Hateful Multimodal Memes. In Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021), 207–214.

[3] Vinay Arora, Eddie Yin-Kwee Ng, Rohan Singh Leekha, Medhavi Darshan, and Arshdeep Singh. 2021. Transfer learning-based approach for detecting COVID-19 ailment in lung CT scan. Comput. Biol. Med. 135, (2021), 104575.

[4] Marina Bulatovic. 2019. The imitation game: The memefication of political discourse. Eur. View 18, 2 (2019), 250–253.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. ArXiv Prepr. ArXiv181004805 (2018).

[6] Priya Dixit. 2022. Memeing the Far-Right: Pepe and the Deplorables. In Race, Popular Culture, and Far-right Extremism in the United States. Springer, 135–172.

[7] Jarod Govers, Philip Feldman, Aaron Dant, and Panos Patros. 2023. Down the Rabbit Hole: Detecting Online Extremism, Radicalisation, and Politicised Hate Speech. ACM Comput. Surv. (2023).

[8] Taranjit Kaur and Tapan Kumar Gandhi. 2019. Automated brain image classification based on VGG-16 and transfer learning. In 2019 International Conference on Information Technology (ICIT), IEEE, 94–98.

[9] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. Adv. Neural Inf. Process. Syst. 33, (2020), 2611–2624.

[10] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, and Siqi Yan. 2020. Captum: A unified and generic model interpretability library for pytorch. ArXiv Prepr. ArXiv200907896 (2020).

[11] Yikuan Li, Hanyin Wang, and Yuan Luo. 2020. A comparison of pre-trained vision-and-language models for multimodal representation learning across medical images and reports. In 2020 IEEE international conference on bioinformatics and biomedicine (BIBM), IEEE, 1999–2004.

[12] Alice E. Marwick and Rebecca Lewis. 2017. Media manipulation and disinformation online. (2017).

[13] Lawankorn Mookdarsanit and Pakpoom Mookdarsanit. 2021. Combating the hate speech in Thai textual memes. Indones. J. Electr. Eng. Comput. Sci. 21, 3 (2021), 1493–1502.

[14] Davide Mulfari, Antonio Celesti, Maria Fazio, Massimo Villari, and Antonio Puliafito. 2016. Using Google Cloud Vision in assistive technology scenarios. In 2016 IEEE symposium on computers and communication (ISCC), IEEE, 214–219.

[15] Joel Penney. 2020. 'It's So Hard Not to be Funny in This Situation': Memes and Humor in US Youth Online Political Expression. Telev. New Media 21, 8 (2020), 791–806.

[16] Zhongang Qi, Saeed Khorram, and Fuxin Li. 2019. Visualizing Deep Networks by Optimizing with Integrated Gradients. In CVPR Workshops.

[17] Yiting Qu, Xinlei He, Shannon Pierson, Michael Backes, Yang Zhang, and Savvas Zannettou. 2022. On the Evolution of (Hateful) Memes by Means of Multimodal Contrastive Learning. ArXiv Prepr. ArXiv221206573 (2022).

[18] Jamie Noelle Smith. 2018. No Laughing Matter: Failures of Satire During the 2016 Presidential Election. (2018).

[19] Zongwei Song. 2019. Is the Spreading of Internet Neologisms Netizen-Driven or Meme-driven? Diachronic and Synchronic Study of Chinese Internet Neologism Tuyang Tusen Po. Theory Pract. Lang. Stud. 9, 11 (2019), 1424–1432.

[20] Xiaocui Yang, Shi Feng, Daling Wang, and Yifei Zhang. 2020. Image-text multimodal emotion classification via multi-view attentional network. IEEE Trans. Multimed. 23, (2020), 4014–4026.

[21] Jiawen Zhu, Roy Ka-Wei Lee, and Wen Haw Chong. 2022. Multimodal zero-shot hateful meme detection. In 14th ACM Web Science Conference 2022, 382–389.

[22] Explained: 'Trads' vs 'Raitas' and the Inner Workings of India's Alt-Right. Retrieved May 1, 2022 from https://thewire.in/communalism/genocide-as-pop-culture-inside-the-hindutva-world-of-trads-and-raitas