9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

Crowdsourcing to Identify Hateful Memes: A Human-in-the-Loop Approach

hateful meme,crowdsourcing,human-in-the-loop,hate speech,social media

Extended Abstract

Image-based social platforms, such as Instagram or Pinterest, mandate the inclusion of non-textual content, namely images and video clips, in every post. This requirement has made visual communication an integral and indispensable element of modern social networks. Among these platforms' various types of image content, memes have emerged as a prevalent and viral format due to their inherent reproducibility, which allows them to be easily shared and remixed by users, and their ability to convey complex ideas or emotions concisely and humorously. Additionally, memes often serve as a form of cultural commentary or critique, allowing users to express their opinions using both images and text. These memes are commonly known as image-with-text memes (IWT-memes). However, the prevalence of memes on image-based social platforms can also have negative consequences. In particular, IWT-memes that include explicit or implicit comments that ridicule other users, especially those with mental or physical disabilities, can spread quickly and widely. These memes can also be manipulated to disseminate malicious information, such as racially, religiously, politically, or sexually offensive content. Despite their potential for harm, few studies have been conducted to address this novel and hybrid form of hate speech. As such, there is a pressing need better to understand the impact of IWT-memes on social networks and develop strategies to mitigate their harmful effects. Previous research has examined using neural-network models to detect and cluster memes on social media [Lee et al., 2021, Yang et al., 2022, Du et al., 2020]. Still, there has been limited exploration of how to identify hateful information. Our paper proposes a method to detect and classify hateful IWT-memes on social media platforms. We leverage the wisdom of crowds to improve accuracy and account for the subjective nature of identifying hateful content.

Methodology

Our identification scheme consists of two classification models implemented in two stages, as illustrated in Figure 1. In the first stage, our model aims to differentiate between IWT-memes and Non-IWT-memes. To achieve this, we split a meme into the image and the textual content. We then construct a classifier using neural networks to distinguish IWT-memes from a collection of images. In the second stage of our study, we aimed to classify IWT-memes into two categories: non-hellish IWT-memes and hellish IWT-memes. The latter category comprises attack, and harmful image content conveyed through suggestive text. Hellish IWT-memes often contain provocative, offensive, and derogatory language or imagery intended to harm or insult a particular individual or group. Such content may include hate speech, racism, sexism, homophobia, and other forms of discrimination. Figure 1 depicts examples of various types of IWT-memes. The non-hellish IWT-meme in Figure 1(a) shows a plane being carried by another plane with the caption "That's how planes are born," and is described humorously without any hateful content. On the other hand, by maliciously altering the spelling of "Handsome," the hellish IWT-meme in Figure 1(b) pokes fun at disabled people. However, identifying suggestive

content like hate speech can be challenging for a machine learning model due to the nuanced and subtle forms it can take. Such content often contains implicit or indirect references to individuals or groups and can vary in their degree of offensiveness and context dependency.

To address these challenges, we propose integrating human judgment into the identification loop. Specifically, we leverage crowdsourcing to incorporate diverse perspectives and expertise into the classification process, making the infrastructure more flexible and scalable [Wu et al., 2021]. By using crowdsourcing, we can also mitigate potential biases that may arise from relying solely on the judgments of a small group of experts. Additionally, by involving many individuals in the identification process, we can increase the speed and efficiency of the workflow.

Furthermore, we utilize paired comparison, a widely used method in crowdsourcing, to quantify the degree of offensiveness and harness of the identified IWT-memes. This approach allows us to compare and rank memes based on their relative level of harm, providing a more fine understanding of the severity of different types of IWT-memes. Using paired comparison also helps ensure that the identification process is more objective and consistent, reducing the influence of individual biases or preferences.

Conlcusions

In conclusion, our study contributes to hate speech detection and classification on social media platforms. By focusing on identifying and classifying hellish IWT-memes, we believe the importance of incorporating human judgment and crowdsourcing into the identification process could improve the accuracy and effectiveness of the classification models. Our proposed approach using paired comparison and crowdsourcing offers a flexible and scalable solution to identifying forms of hate speech in non-textual content. It allows for incorporating human input and judgment, which is crucial in recognizing the context and complexity of such content.

In the future, ongoing research and monitoring of hate speech on social media platforms, particularly in non-textual content like images and video clips, will be crucial. By adopting our mechanism, we aim to better understand how such memes impact victims on social media. This enables us to develop more effective strategies to mitigate their impact and promote a more inclusive and peaceful online environment. We hope our findings inspire further research in this rapidly evolving field and ultimately contribute to creating a safer and more positive online community for all users.

References

[Du et al., 2020] Du, Y., Masood, M. A., and Joseph, K. (2020). Understanding visual memes: An empirical analysis of text superimposed on memes shared on twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):153–164.

[Lee et al., 2021] Lee, R. K.-W., Cao, R., Fan, Z., Jiang, J., and Chong, W.-H. (2021). Disentangling hate in online memes. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5138–5147.

[Wu et al., 2021] Wu, A., Xie, L., Lee, B., Wang, Y., Cui, W., and Qu, H. (2021). Learning to automate chart layout configurations using crowdsourced paired comparison. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

[Yang et al., 2022] Yang, C., Zhu, F., Liu, G., Han, J., and Hu, S. (2022). Multimodal hate speech detection via cross-domain knowledge transfer. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4505–4514.

(a) Non-hellish IWT-memes

(b) Hellish IWT-memes

Figure 1: Examples of different types of IWT-memes.

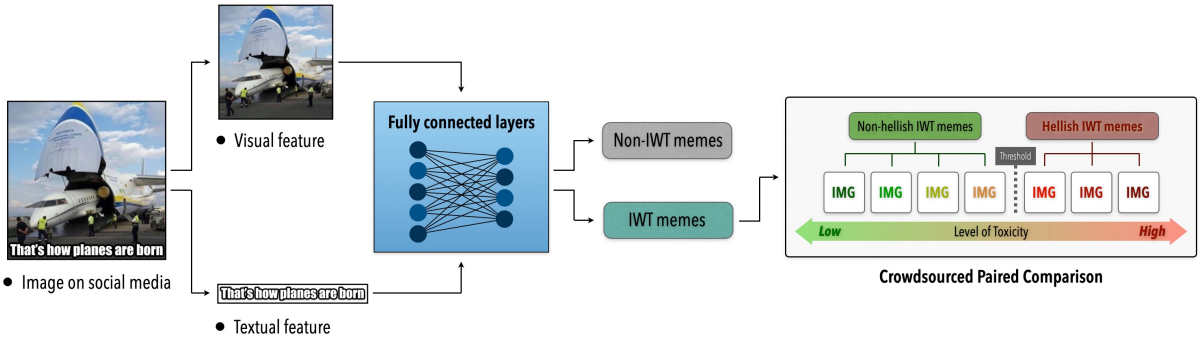

Figure 2: Workflow of our identification scheme