Detecting Harmful Image With Text Memes: A Multimodal Analysis

[1]*Zih-Rong Lin* (林資融) [1]*Chiou-Shann Fuh* (傅楸善)

[1]Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan
*E-mail: r11944078@csie.ntu.edu.tw fuh@csie.ntu.edu.tw

ABSTRACT

Social network platforms have witnessed the increasing importance of visual communication. Images and multimedia content play a crucial role in engaging users, and among these, Image With Text (IWT) memes have emerged as a popular form of communication. However, the prevalence of harmful or offensive content in these memes can have a negative impact on society. In this paper, we propose a multimodal approach for identifying hellish IWT memes. Our approach leverages features extracted from IWT memes, including visual, text, and facial features, to accurately identify these harmful memes. We compare the effectiveness of these features in the classification process and provide insights into which features are most valuable for identifying these memes. Furthermore, we conduct experiments on a dataset of IWT memes to evaluate the performance of our approach and demonstrate that it outperforms existing methods for identifying hellish IWT memes. Our findings suggest that a multimodal approach that incorporates visual, text, and facial features can effectively detect and classify hellish IWT memes.

Keywords: hellish IWT memes

1. INTRODUCTION

Based on our observations, it is apparent that IWT has attracted more participants in the past year, with IWT memes being one of the most popular features. The findings presented in Figure 1 demonstrate that employing the IWT meme can result in a substantially higher level of engagement, indicating that it may be an effective strategy for conveying information. This popularity allows IWT memes to spread rapidly and effectively convey information. Unfortunately, some individuals exploit IWT memes to disseminate harmful information to users, particularly those with mental or physical disabilities. Furthermore, these types of IWT memes often contain offensive content related to race, religion, politics, or sexual orientation.

Despite the existence of numerous papers focusing on classifying IWT memes and non-IWT memes using text features in images and visual content, which have shown promising performance, there has been limited research dedicated to identifying IWT memes that contain hateful content. Hence, we decided to apply the aforementioned features to detect hellish IWT memes. During our investigation, we discovered that there is still potential for improvement in terms of accuracy. Therefore, our primary focus lies in distinguishing between normal IWT memes and hellish IWT memes, with the aim of enhancing the classification performance.

Beskow et al. [1] have shown that a multi-modal model achieves higher accuracy than a uni-modal model for classifying IWT memes and non-IWT memes, so we adopt a multi-modal model to implement the classification of memes in our research. As we know, IWTs contain not only image features but also text and other relevant information. Therefore, we utilize features extracted from visuals, text, and faces enhancing our performance in identifying hellish IWT memes. Our research shows that the multimodal model, which incorporates visual, text, and facial features, achieves an

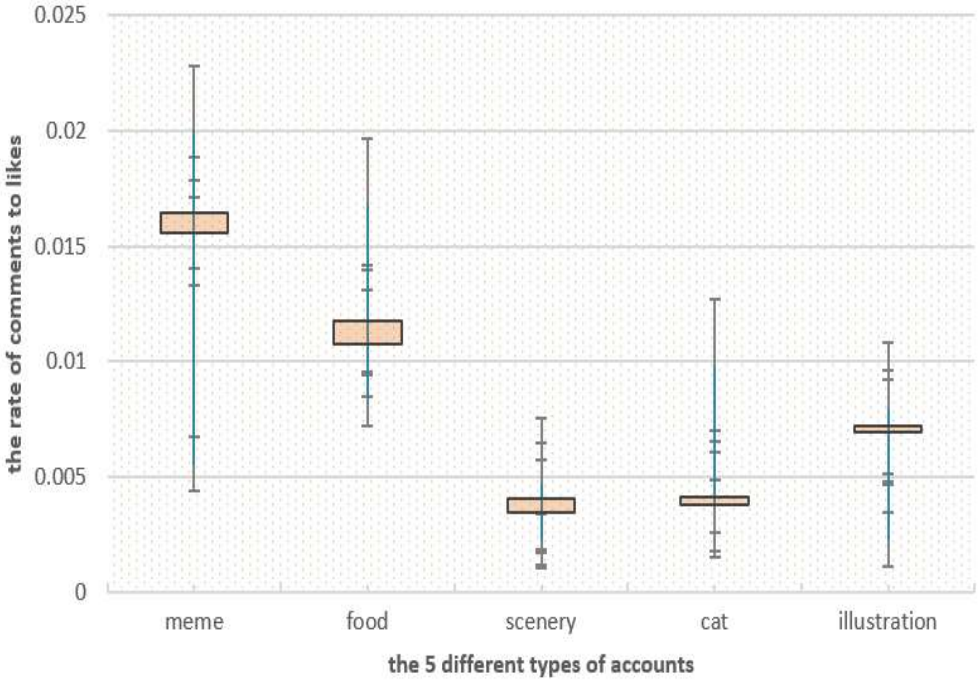

Fig. 1. The box plot about the rate of comments to likes on 100 pictures at 5 different types of account

accuracy of 73.7%, outperforming the unimodal model that uses only visual features with an accuracy of 72% and the multimodal model that uses only visual and text features with an accuracy of 73.1%. Additionally, based on our observations, it is obvious that the strategic incorporation of facial features has resulted in a significant boost in the recall of our model. This positive outcome has notably improved its capability to detect and retrieve the elusive hellish IWT meme with greater efficiency.

2. RELATED WORK

In this section, we will review prior research on the analysis of memes.

2.1. Memes

Currently, there is a significant amount of research focused on analyzing memes, regardless of their format, such as images with text, videos, and so on. Xie et al. [2] conducted a study using YouTube to identify frequently reposted short video segments that they refer to as “video memes”. They created a graph of people and content to model interactions between these video memes. However, analyzing image memes is more complex as it involves first classifying an image as a meme or not. To address this, Dubey et al. [3] developed the Memesequencer model, which separates the underlying image template from additional text and image manipulation in a meme image. The model creates a meme embedding by combining image and text features using deep learning techniques, with the best model combining ResNet18[4] with SkipThought text features. After generating an embedding, the authors construct an evolutionary tree using a phylogenetic approach to represent the relationships between different memes.

2.2. The Digital Footprint of Memes

Additionally, The digital footprint left by Internet memes enables researchers to analyze

how memes spread through various networks. Bauckhage et al. [5] investigated the temporal models of fads by analyzing Internet memes and approximating interest in a particular meme using Google Trends. Leskovec et al.[6] utilized memes and phrases extracted from news and blogs to track and examine the dynamics of the news cycle. This research mapped the evolution of text-based memes in the blogosphere and news cycle, allowing for a deeper understanding of how memes spread through these networks.

2.3. Image-based Memes and IWT Memes

Previous research on image analysis can be categorized into two types: image-based memes and image-with-text memes. In the case of image-based memes, Bharati et al. [7] proposed a method called "Provenance analysis" to detect memes and investigate their intended purpose and motivation. On the other hand, Zannettou et al. [8] conducted a comprehensive study on the prevalence of anti-Semitism in alt-right web communities, quantifying the growth of its online dissemination.

In the case of image-with-text memes, Beskow et al. [1] developed a multi-modal deep learning model to differentiate between memes and non-memes. They compared their approach with single-modal methods for categorizing political memes online and utilized graph learning to construct evolutionary trees of memes. Additionally, as we mentioned before, Dubey et al. [3] used pre-trained deep neural networks and optical character recognition (OCR) [9] to extract meme features, which proved useful in tasks such as image clustering, retrieval, topic prediction, and virality prediction. Similarly, Du et al. [10] utilized a similar technique to identify themes in image-with-text memes and found that 30% of identifiable themes were related to politics and text features.

However, our research aims at finding the hellish IWT memes and trying to improve the classification accuracy between non-hellish IWT memes and hellish IWT memes with multimodal models adding face features, and comparing which combination of features is better.

(a) normal IWT meme

(b) hellish IWT meme

Fig. 2. Examples of different kinds of IWT.

3. METHOD

In this section, we will present a more precise definition of IWT memes and hellish IWT memes. Additionally, we will provide a detailed description of the dataset employed in this study and explicate how we utilized the features extracted from the IWT meme to enhance performance. Figure 1 is the experimental design of our model.

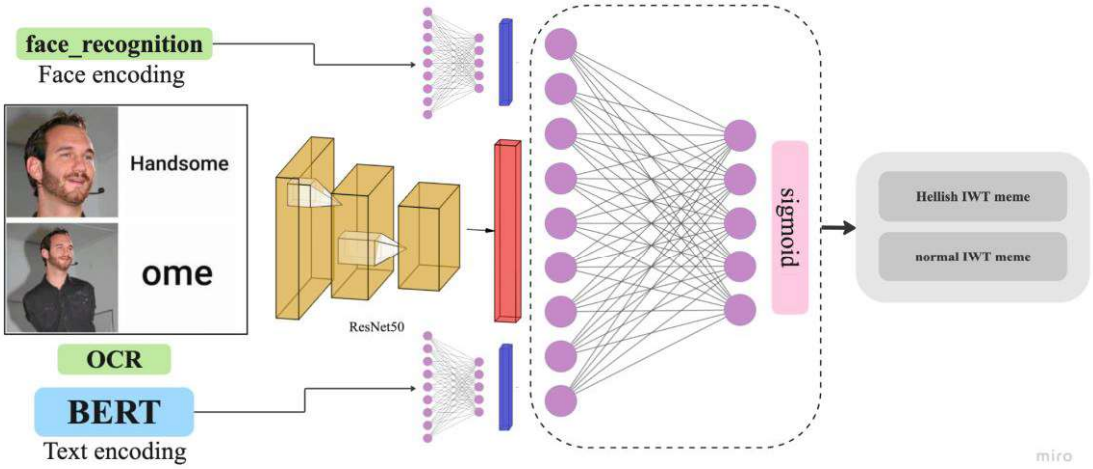

Fig. 3. Experimental design

3.1. Identification of Data Types

Our goal in the following paragraphs is to offer concise yet comprehensive definitions and detailed explanations for both normal IWT memes and hellish IWT memes, which will allow readers to gain a more profound understanding of these concepts and the potential consequences they may have.

Normal IWT memes

According to the Oxford English Dictionary, a meme is defined as "An image, a video, a piece of text, etc. that is passed very quickly from one internet user to another, often with slight changes that make it humorous." Therefore, IWT memes usually involve various visual materials such as movie scenes, celebrity photos, and news items, which feature a superimposed text that expresses a particular situation or provides a humorous or ironic commentary on the image. Often, memes reflect current events or popular culture, making them a quick and relatable way for people to engage with and comment on the world around them. Figure 2(a) is an example of an IWT meme.

Hellish IWT memes

Within the realm of IWT memes, the most widely disseminated are those with a black humor theme, commonly referred to as hellish IWT memes. These memes often incorporate tragic, discriminatory, and disease-related content, and are frequently used to taunt or disparage specific groups, sometimes even crossing ethical boundaries. While these memes can be controversial and offensive to some, they are often shared and enjoyed by those who appreciate dark humor and satire, highlighting the subjective nature of humor. Despite their potentially harmful nature and the criticism they receive, hellish IWT memes have become a popular and influential form of online expression, shaping the way people engage with and respond to social and political issues. Figure 2(b) is an example of a hellish IWT meme.

3.2. Dataset

To assess the performance of our model, we conducted experiments on a dataset of image-with-text (IWT) memes crawled by Zih-Rong Lin et al. This dataset consists of Chinese-

language IWT memes collected from Instagram and labeled as either "normal" or "hellish" based on human categorization. In cases where the meaning of an IWT meme was ambiguous and difficult to classify, three individuals used a voting system to reach a consensus. There are 785 normal IWT memes and 638 hellish IWT memes in this dataset. The training and validation data were divided into an 8:2 ratio, and we utilized cross-validation techniques to calculate our accuracy.

3.3. Memes Classification Models

Du et al. [10] propose that a classifier model consists of three primary components: a visual feature extractor, a textual feature extractor, and a meme classifier. In this paper, we have extended the model to include a face feature extractor to detect our hellish IWT meme, making it easier for the model to identify this specific meme. The following explains how each feature extractor contributes to detecting the hellish IWT meme and how we used them.

Visual feature

Given our specific focus on image-based meme detection, it is highly advantageous to employ Convolutional Neural Networks (CNNs) [11] as they have emerged as the most popular and effective models for visual learning tasks. Consequently, adopting a CNN-based approach is a natural choice for our work. To this end, we utilized the ResNet50 [4] architecture, renowned for its exceptional performance in image recognition tasks, to extract pertinent visual features. This choice not only ensures effective feature extraction but also lays a solid foundation for further advancements in our meme detection framework.

Text feature

To extract textual features, we employed a robust and widely adopted approach. Initially, we utilized the highly capable Optical Character Recognition (OCR) engine, Tesseract [9], to accurately extract characters from images. By leveraging its advanced algorithms, we effectively obtained the textual information present within the meme images. Next, we employed Bidirectional Encoder Representations from Transformers (BERT) [12], a cutting-edge transformer-based model, to translate these characters into high-dimensional vectors. BERT has demonstrated remarkable success in various natural language processing tasks and possesses a deep understanding of contextual information. By leveraging the power of BERT, we were able to capture the semantic meaning and contextual relevance of the textual content extracted from the memes. To further refine our textual feature representation, we calculated the element-wise average of these BERT vectors. This step enables us to obtain a condensed yet comprehensive representation that encapsulates the salient information within the textual content of the memes.

Face feature

We are aware that there are numerous offensive Internet memes that target specific individuals or groups, including those based on their race, disability, and other characteristics. These memes often highlight facial features, making the extraction of facial features an important aspect of analyzing them. To accomplish this, we leverage the power of the widely adopted open-source face detection software package called face_recognition [13]. By employing this cutting-edge library, we can accurately detect faces within the meme images and obtain face encoding vectors for each detected face. These face encoding vectors serve as invaluable inputs to our classification models, as they encapsulate essential facial characteristics and nuances. By utilizing this rich facial information, our models gain a comprehensive understanding of the unique facial expressions and features present in the memes. This enables us to effectively discern and interpret the visual cues embedded in the memes, ultimately enhancing the accuracy and reliability of our meme detection system.

Table 1. Result.

	Accuracy	Precision	Recall	F1-score
Vision	0.722	0.688	0.697	0.690
Vision + Text	0.731	0.697	0.711	0.703
Vision + Text + Face	**0.737**	**0.701**	**0.731**	**0.713**

Fully Connected Layer

In the classification phase, we adopted a carefully designed approach to achieve optimal results. Initially, we modified the pre-trained ResNet50 network by removing its last fully connected layer. This adjustment allowed us to extract and retain the crucial visual features essential for meme classification. Subsequently, we introduced a new fully connected network that combined the text feature vector and face feature vector. These vectors were concatenated to capture the textual and facial aspects of the memes comprehensively. To ensure effective classification, we added a sigmoid layer at the end to produce the desired probability outputs. To train our model effectively, we employed a two-step process. Initially, we focused solely on training the CNN-based model to obtain excellent pre-trained weights. This step allowed us to benefit from the powerful feature extraction capabilities of the pre-trained network. Next, we initialized our model, incorporating the newly added layer, with the pre-trained weights from the CNN. We intentionally froze the weights of the CNN portion to preserve the knowledge learned during the initial training phase. By doing so, we directed our attention towards training the fully connected layer specifically. This strategic approach enabled us to fine-tune and optimize the model's performance, resulting in improved classification results. By combining these steps, we achieved a comprehensive and effective classification framework that leverages the strengths of both the pre-trained CNN and the newly added fully connected layer. This methodology enhances our model's ability to detect hellish IWT memes, leveraging both visual, textual and facial features to provide superior results accurately.

(a) hellish IWT meme

(b) normal IWT meme

Fig. 4. Examples of misclassified IWT meme.

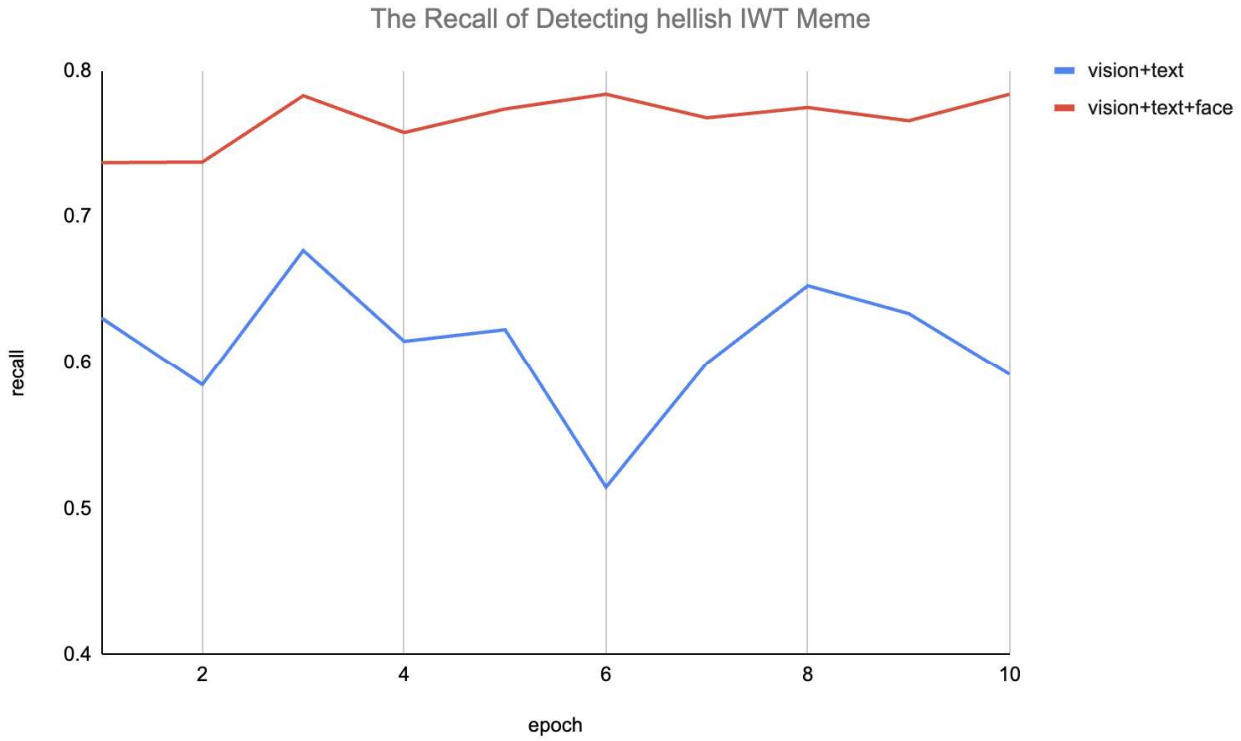

Fig. 5. The recall of detecting hellish IWT meme

4. RESULTS

Based on the findings presented in Table 1 and Figure 5, it is evident that the integration of face features into our model significantly enhances its performance, particularly in terms of recall. The results indicate that while the overall accuracy (Vision + Text + Face) improves by a modest 0.6% compared to the Vision + Text approach, there is a substantial increase of 2% in recall. This improvement suggests that our model is better equipped to successfully detect hellish IWT memes, thereby facilitating the accomplishment of the task. Furthermore, these results highlight the crucial role played by the inclusion of face features in effectively identifying the hellish IWT meme.

However, our model still struggles to detect certain types of hellish memes. When examining the Figure 4(a), we can observe that these hellish IWT memes contain complex meanings and historical references, making it challenging for our model to accurately identify them. Additionally, the Figure 4(b) is misclassified as hellish memes, possibly due to the presence of a black man's face, which leads to misclassification by our model. These observations indicate that our model still has limitations in detecting certain types of hellish memes. In the future, further improvements are needed to address the challenges posed by memes with intricate meanings and diverse visual content, such as historical references and facial features.

5. CONCLUSION

In conclusion, this study showcases the effectiveness of integrating face features in improving the detection of hellish IWT memes. The results demonstrate the model's enhanced ability to identify such memes, with a significant increase in recall. However, the study also highlights the existing limitations, particularly in detecting memes with intricate meanings and diverse visual content. Future efforts should be directed towards refining the algorithms and incorporating more advanced techniques to overcome these challenges and achieve greater accuracy in identifying hellish IWT memes. The findings of this research contribute to the growing body of knowledge

in meme classification and pave the way for further advancements in meme detection systems.

6. REFERENCES

[1] Beskow, David M., Sumeet Kumar, and Kathleen M. Carley. "The evolution of political memes: Detecting and characterizing internet memes with multi-modal deep learning." Information Processing & Management 57.2 (2020): 102170.

[2] Xie, Lexing, et al. "Visual memes in social media: tracking real-world news in youtube videos." Proceedings of the 19th ACM international conference on Multimedia. 2011.

[3] Dubey, Abhimanyu, et al. "Memesequencer: Sparse matching for embedding image macros." Proceedings of the 2018 world wide web conference. 2018.

[4] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016

[5] Bauckhage, Christian, Kristian Kersting, and Fabian Hadiji. "Mathematical models of fads explain the temporal dynamics of internet memes." Proceedings of the International AAAI Conference on Web and Social Media. Vol. 7. No. 1. 2013.

[6] Leskovec, Jure, Lars Backstrom, and Jon Kleinberg. "Meme-tracking and the dynamics of the news cycle." Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. 2009.

[7] Bharati, Aparna, et al. "Beyond pixels: Image provenance analysis leveraging metadata." 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2019.

[8] Zannettou, Savvas, et al. "On the origins of memes by means of fringe web communities." Proceedings of the internet measurement conference 2018. 2018.

[9] Smith, Ray. "An overview of the Tesseract OCR engine." Ninth international conference on document analysis and recognition (ICDAR 2007). Vol. 2. IEEE, 2007.

[10] Du, Yuhao, Muhammad Aamir Masood, and Kenneth Joseph. "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. Vol. 14. 2020.

[11] Albawi, Saad, Tareq Abed Mohammed, and Saad Al-Zawi. "Understanding of a convolutional neural network." 2017 international conference on engineering and technology (ICET). Ieee, 2017.

[12] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[13] Geitgey, Adam. "Face recognition documentation." Release 1.3 (2019): 3-37.