

Abstract

With the widespread growth of social media platforms, human-only moderation has become impractical due to the volume of new content continuously released. The goal of multimodal hateful meme analysis is to create computational models that can effectively tell whether memes, images with text captions that are usually intended to be humorous, are hateful and should be removed. These models can be split into two groups: early fusion, and late fusion. In this paper, we evaluate early fusion and late fusion models to see which is more practical for a real-world setting. Both of our models were made with ResNet-50 and DistilBERT as base models. In terms of performance, our early fusion model had a 0.9413 ROC-AUC, while our late fusion model had a 0.9438 ROC-AUC. However, the early fusion model outperformed the late fusion odel on accuracy, precision, recall, and F1 Score, and better balanced the modalities. Our late fusion model had a high reliance on the visual modality, while our early fusion model was well-balanced but leaned towards the text modality. We conclude by stating that early fusion is more balanced across the different modalities than late fusion, and therefore is a more practical choice for hateful meme analysis.

Introduction

Social Media has become an integral part of everyday life, and today over 90% of adolescents in the United States of America use platforms like Facebook and Instagram daily. Billions of posts are exchanged daily between users all over the world, and one of the most commonly used post formats is the meme: it consists of an image and a text caption and uses numerous inside jokes. Although many memes are lighthearted and non-offensive, social media platforms have encountered difficulties with hateful memes “toxifying” communities, and removing them is an important priority. Although human moderators are more accurate, at the scale modern social media operates at, it is impossible to hire enough workers. Therefore, companies like Meta have tried to create machine learning models that can flag these memes; Although hateful memes are hard to classify with solely AI, they can act as a “first filter.” In this project, I explore two foundational machine learning paradigms, early and late fusion, and compares them in the context of hateful meme analysis. Early fusion merges the image and text from the start, while Late fusion processes them individually and then combines them afterwards.

Methods

In order to accurately compare early fusion and late fusion, both need to use similar base models, the same dataset, and be trained with a similar capacity. Therefore, larger models aren’t really the key here. In order to preserve accuracy while maintaining reasonable training time, I used a ResNet-50 CNN and a DistilBert Transformer. For the Early Fusion model, we chose to feed the outputs of the ResNet model into the input of the DistilBERT model. I did this via accessing the last hidden state and passing it through a simple fully-connected layer, compressing it into a format that could be seamlessly merged with the text tokens. For the Late Fusion model, we merged the outputs of the ResNet and DistilBERT models into a multi-layer perceptron that gradually shrunk into one output node, the hatefulness. We did this by taking the ResNet pooler output and the DistilBERT output token and using a small dropout at each step; randomly setting nodes to zero to make sure the model doesn’t rely too much on one node. For both models, we applied Binary Cross-Entropy loss, as it is the standard loss function to use when there are only two possible outputs.

Results

When both modalities were used, the late fusion model had 0.8988 Accuracy, 0.8717 Precision, 0.8454 Recall, 0.8583 F1-Score, and 0.9438 ROC-AUC. The early fusion model had 0.9086 Accuracy, 0.8905 Precision, 0.8529 Recall, 0.8713 F1-Score, and 0.9413 ROC-AUC. When only text was used, the late fusion model only outputted zero, while the early fusion model had 0.7278 Accuracyy, 0.6167 Precision, 0.6594 Recall, 0.6374 F1-Score, and 0.7658 ROC-AUC. With only the image, the late fusion model had 0.7555 Accuracy, 0.7419 Precision, 0.5005 Recall, 0.5978 F1 Score, and 0.8140 ROC-AUC. The Early Fusion model had 0.6886 Accuracy, 0.6356 Precision, 0.3318 Recall, 0.4360 F1-Score, and 0.7022 ROC-AUC.

Discussion

When both text and image are provided, both models perform similarly. The Early Fusion model outperforms the Late Fusion model, but they are very close. However, these results are contradictory with another study that used a different dataset. I think this is because of the different dataset and model usage. When only text is used, early fusion performs much better, since late fusion only outputs zero. One potential conclusion from this is that the late fusion model relies more on the visual modality. Indeed, the scores with the visual modality only are even better than the Early Fusion model, and this shows a higher reliance on images as context in late fusion models. Conversely, the Early Fusion model performed better than if it only had the image, demonstrating that it relied more on text.

Conclusion

For this project, I measured the capabilities of early fusion models compared to late fusion models in the context of multimodal meme analysis, and I used the Facebook Hateful Memes dataset to train my models. Overall, early fusion performs similarly to late fusion. However, late fusion has a heavy reliance on images, while early fusion models perform well while missing either modality. This suggests that early fusion is more well-balanced, while late fusion relies more on the image modality. In the real world, this may mean that the early fusion model will perform better than the late fusion model. Late fusion might also be deceived by hateful speech paired with an inconspicuous image. Overall, early fusion is a better choice than late fusion in the real world for multimodal hate meme analysis.

References

[1] Faseela Abdullakutty and Usman Naseem. [4] Yogesh Gajula. SENTIMENT-AWARE “Decoding Memes: A Comprehensive Analysis of Late and Early Fusion Models for E-commerce: A Review from a Natural Language Processing Perspective. Tech. rep. University of ACM Web Conference, Association for Computing Machinery, Inc, May 2024, pp. 1681–1688. isbn: 9798400701726. doi: 10.1145/3589335.3652504. [2] Rui Cao et al. “Prompting for Multimodal Walkthrough Analysis”. In: JMIR Human Factors 12 (Jan. 2025), issn: 2292-9495. doi: 10.2196/69817. url: https://www.sciencedirect.com/journal/science/article/pii/S2292949525001300. [3] Cech N. Dang, Mar’ia N. Moreno-García and Fernando De la Prieta. “An approach to integrating sentiment analysis into recommender systems”. In: Sensors 21 (16 Aug. 2021), issn: 14248220. doi: 10.3390/s21165666. [8] Douwe Kiela et al. “The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes”. In: (Apr. 2021). url: http://arxiv.org/abs/2005.04790. [9] Hannah Rose Kirk et al. “Memes in the Wild: Assessing the Generalizability of the Hateful Memes Challenge Dataset”. In: (July 2021). url: http://arxiv.org/abs/2107.04313. [10] Dita Ayu Lestari et al. “Meme Culture: A Study of Humor and Satire in Digital Media”. In: International Journal of Advanced Multidisciplinary Research and Studies 4 (4 July 2024), pp. 134–140. doi: 10.62225/2530409.2024.4.4.3013. [11] Victor Sanh et al. “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter”. In: (Nov. 2020). url: http://arxiv.org/abs/1910.01108. [12] Carol F. Scott et al. “Time spent online: Latent profile analyses of emerging adults’ social media use”. In: Computers in Human Behavior 75 (Oct. 2017), pp. 311–319. issn: 07475632. doi: 10.1016/j.chb.2017.05.026. [13] Bryan Zhao et al. “A Review of Vision-Language Models and their Performance on the Hateful Memes Challenge”. In: (May 2023). url: http://arxiv.org/abs/2305.06159. [14] Liangliang Zhao, Shuyi Huang, and Junsong Li. “Multimodal Data Fusion Research Review: From Traditional Methods to Unsupervised Learning Methods”. In: Proceedings of 2025 2nd Symposium on Big Data, Neural Networks, and Deep Learning. BDNNDL 2025. Association for Computing Machinery, Inc, Jan. 2026, pp. 249–254. isbn: 9798400720185. doi: 10.1145/3784013.3784052. [15] Ron Zhu. “Enhance Multimodal Transformer With External Label And In-Domain Pretrain: Hateful Meme Challenge Winning Solution”. In: (Dec. 2020). url: http://arxiv.org/abs/2012.08290.



Hateful Meme Examples

```
... classifier.1.weight | UNEXPECTED | |

Notes:
- UNEXPECTED: can be ignored when loading from different task/architecture; not ok if you
Down Loading...
From (original): https://drive.google.com/uc?id=1UL9GBMnkoox9mpFbaUn-cpxoPe8hg-s-
From (redirected): https://drive.google.com/uc?id=1UL9GBMnkoox9mpFbaUn-cpxoPe8hg-s-&confirm
To: /content/lateFusion20.pth
100%|██████████| 372M/372M [00:04<00:00, 90.8MB/s]
<All keys matched successfully>
```

Load Early Fusion Model

Show code

Train Model

Show code

Source code for the models