Meme or Menace? Dissecting Hateful Content in the Age of Social Media

Akansha Lalwani
Data Science
A59019733

Mayank Sharma
Electrical and Computer Engineering
A59019439

Abstract

Hateful meme classification is an emerging and critical issue in the intersection of computer vision and natural language processing. It targets the automated identification and categorization of memes conveying offensive or harmful content. Despite the power of traditional text-based and image-based analysis methods, they still lack a profound understanding of the semantic relationship between textual and visual meme content. This report delves into the exploration and application of deep learning techniques including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-based models like BERT and ViT to handle this problem. We present our project motivation, problem definition, data description, methodologies, results, and analysis.

1 Problem Definition

In recent years, the widespread use of social media has led to an exponential increase in user-generated content, including the rapid dissemination of internet memes. While many memes are light-hearted and humorous, a concerning number of them contain hateful, offensive, or harmful messages. These hateful memes contribute to the spread of negative sentiments, cyberbullying, and discrimination, posing a serious threat to the well-being and mental health of online users.

1.1 Motivation

The motivation for addressing the problem of hateful meme classification stems from multiple factors, which can be categorized into three primary areas:

- **Societal Impact:** Hateful memes perpetuate stereotypes and prejudice, causing harm to targeted individuals or groups. Accurate classification helps mitigate their impact, protect vulnerable populations, and foster inclusive online spaces promoting tolerance and understanding.
- **Psychological Well Being:** The spread of hateful memes negatively affects mental health, causing anxiety, depression, and even suicidal thoughts among targeted individuals. Addressing hateful meme classification promotes healthier online experiences and reduces psychological harm.
- **Technological Advancements:** With increasing user-generated content, human moderation becomes difficult. Advanced machine learning models for hateful meme classification offer scalable solutions for content moderation, improving detection and response efficiency, and contributing to advancements in natural language processing and computer vision.

1.2 Key parts of the problem

The key parts of the problem of hateful meme classification can be broken down into the following aspects:

- **Defining Hateful Memes:** Establishing a clear and comprehensive definition of hateful memes is essential for effectively addressing the problem. This involves understanding the various forms that hateful content can take, such as hate speech, offensive language, derogatory images, or a combination of these elements. Additionally, it is crucial to consider the context and cultural factors that may influence the interpretation of a meme's content as hateful or non-hateful.
- **Multimodal Feature Engineering:** As hateful memes typically integrate text and images, it is crucial to develop sophisticated feature engineering techniques capable of capturing the interplay between these modalities. This includes identifying distinctive visual and textual features, as well as their relationships, that can effectively differentiate hateful memes from non-hateful ones.
- **Dataset Creation and Annotation:** Curating a large, diverse, and representative dataset of memes is essential for training and evaluating hateful meme classification models. The dataset should include a wide range of content, with varying degrees of offensiveness and complexity, to ensure the model's generalizability. The process of annotating the dataset with labels for hateful and non-hateful content requires careful consideration of the definition of hateful memes and the potential biases that may be introduced during the annotation process.
- **Model Evaluation and Performance Metrics:** Selecting appropriate performance metrics for evaluating hateful meme classification models is crucial to determine the effectiveness of the proposed solutions. These metrics should account for the specific challenges associated with the problem, such as the imbalanced nature of the dataset or the potential for false positives and false negatives. Furthermore, it is essential to conduct a thorough analysis of the model's performance, identifying potential weaknesses and areas for improvement.

1.3 Understanding of the problem:

The main challenge in hateful meme classification lies in understanding the complex interplay between the text and visual elements present in memes, which often work together to convey meaning, humor, or a particular sentiment. This necessitates the development of models that can effectively understand and process both modalities, taking into account their interdependence.

Memes often use humor, sarcasm, or irony to convey their message. This makes it challenging to determine whether a meme is genuinely hateful or simply a benign joke. Properly classifying such content requires a deep understanding of context, nuances, and cultural references.

The number of non-hateful memes usually far exceeds the number of hateful ones, leading to an imbalance in training data. This can result in biased models that struggle to accurately classify hateful content, as they may be overfitting to the more common non-hateful examples.

Memes are constantly changing, with new templates, references, and contexts emerging all the time. This requires models that can quickly adapt and stay up-to-date with the latest trends to effectively identify and classify novel hateful content.

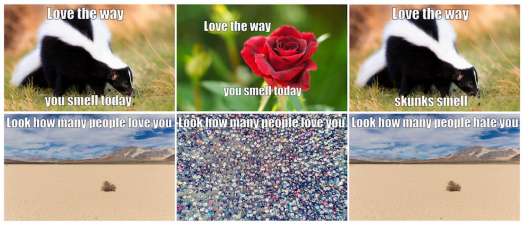
Love the way
you smell today
Love the way
you smell today
Love the way
skunks smell
Look how many people love you
Look how many people love you
Look how many people hate you

Figure 1: Multimodal memes and benign confounders. Hateful memes (left), benign image confounders (middle) and benign text confounders (right).

1.4 Evaluation Metrics

The primary metric used for evaluating the performance of our model in this binary classification task is the Area Under the Receiver Operating Characteristic curve (AUROC). The Receiver Operating Characteristic (ROC) curve is a plot of the True Positive Rate (TPR) against the False Positive Rate (FPR) at various classification thresholds. AUROC can be calculated as follows:

$$AUROC = \int_{-\infty}^{+\infty} TPR(T)\, FPR'(T)\, dT \tag{1}$$

AUROC effectively measures the probability that the model will rank a randomly chosen positive instance higher than a randomly chosen negative instance. The goal is to maximize AUROC.

The secondary metric used for evaluation is accuracy, which calculates the proportion of instances where the predicted class matches the actual class in the test set. Accuracy can be calculated as follows:

$$Accuracy = \frac{1}{N} \sum_{i=0}^{N} I(y_i = \hat{y}_i) \tag{2}$$

where y_i is the actual class, $\hat{y}_i$ is the predicted class, I is the indicator function, and N is the total number of instances.

Ideally, our model should strive to maximize both AUROC and accuracy, to ensure both effective ranking of positive instances and correct classification of all instances.

2 Related Works

The development of machine learning algorithms for text-based hate speech detection has provided a foundation for subsequent work in multi-modal hate speech detection. Schmidt and Wiegand (2017) [1] developed a lexicon-based approach to identify hate speech in social media text, while Davidson et al. (2017) [2] used supervised learning methods to distinguish between hate speech, offensive language, and neither, achieving an F1-score of 0.93.

With the advent of deep learning, researchers started to apply these techniques to hate speech detection. Badjatiya et al. (2017) explored the use of various deep learning models for hate speech detection, finding that Long Short-Term Memory (LSTM) networks outperformed other models. Zhang et al. (2018) used a combination of Convolutional Neural Networks (CNN) and LSTM for hate speech detection, achieving a higher F1-score than models that used CNN or LSTM alone.

Sophisticated multi-modal learning models have been used for hateful meme classification. For instance, Zhou et al. (2020)[3] proposed a model that integrates vision-and-language understanding within a unified Transformer framework, achieving improved performance on the hateful meme detection task.

Recently, there have been significant advancements in this field. In the work titled "Hate-CLIPper: Multimodal Hateful Meme Classification based on Cross-modal Interaction of CLIP Features", the authors proposed an architecture that explicitly models the cross-modal interactions between image and text representations obtained using Contrastive Language-Image Pre-training (CLIP) encoders. This architecture, coupled with a feature interaction matrix (FIM), achieved state-of-the-art performance on the Hateful Memes Challenge (HMC) dataset with an AUROC of 85.8, even surpassing human performance.

Another recent advancement is the introduction of PromptHate in "Prompting for Multimodal Hateful Meme Classification". This model uses prompts to exploit the implicit knowledge in the pre-trained RoBERTa language model for hateful meme classification. The results showed that PromptHate achieved a high AUC of 90.96, outperforming state-of-the-art baselines on the hateful meme classification task.

Lastly, the paper "On Explaining Multimodal Hateful Meme Detection Models" provided insights into what pre-trained visual-linguistic models learn during the hateful meme classification task. The authors found that the image modality contributes more to the hateful meme classification task and that the models can perform visual-text slurs grounding to a certain extent. However, they also observed that these models have acquired biases, which resulted in false-positive predictions.

Despite these advancements, hateful meme classification remains a significant challenge due to the complexity of multimodal data and the subtlety of hate speech. Understanding the semantics of a meme often requires a deep understanding of both the image and text, as well as the cultural context in which the meme is used. Moreover, hate speech can often be subtle and indirect, making it difficult for machine learning models to detect reliably.

3 Method

3.1 Baseline: Sentence Transformer

The Sentence Transformer[4], introduced by Reimers and Gurevych in 2019, serves as an effective model for text classification tasks in general. It leverages transformer architectures for the generation of meaningful sentence embeddings.

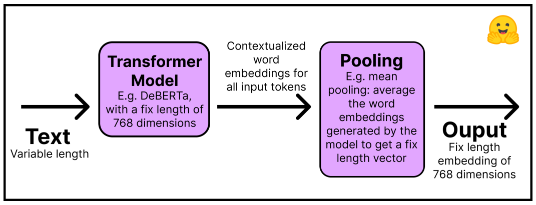
Text
Variable length
Transformer Model
E.g. DeBERTa, with a fix length of 768 dimensions
Contextualized word embeddings for all input tokens
Pooling
E.g. mean pooling: average the word embeddings generated by the model to get a fix length vector
Ouput
Fix length embedding of 768 dimensions

Figure 2: Sentence Transformer

1. **Transformer Model:** At the heart of the Sentence Transformer is a transformer model, typically a BERT[5] or RoBERTa model[6], which generates contextual embeddings for each token in a sentence. The transformer takes a sequence of input word embeddings, applies multiple layers of self-attention and position-wise feed-forward neural networks, and outputs a sequence of contextualized token embeddings.
2. **Pooling Operation:** Given the contextual embeddings for each token, a pooling operation is applied to convert these variable-length sequences into fixed-size sentence embeddings. Common pooling strategies include mean-pooling (averaging all token embeddings), max-pooling (taking the maximum value over each dimension across all tokens), and CLS-pooling (using the special [CLS] token's embedding).
3. **Training:** Sentence Transformers are usually trained using a siamese or triplet network architecture, in which the model learns to minimize the distance between semantically similar sentences while maximizing the distance between semantically dissimilar sentences. This contrastive loss encourages the generation of high-quality sentence embeddings that preserve semantic meaning.

The Sentence Transformer, with its ability to generate semantically meaningful sentence embeddings, can play a crucial role in text-based meme classification tasks. By converting complex sentence structures into a fixed-size vector, it allows for simple and efficient comparison of semantic similarity, making it an effective tool for text classification tasks.

3.2 Baseline: Vision Transformer (ViT)

The Vision Transformer (ViT) introduced a paradigm shift in the computer vision domain, borrowing heavily from the transformers primarily used in Natural Language Processing (NLP). Here is a detailed look at the architecture of the Vision Transformer.

1. **Image Tokenization:** In ViT, an image is treated as a sequence of patches, similar to treating a sentence as a sequence of words in NLP tasks. The input image is divided into a grid of non-overlapping patches, each of the same size (e.g., 16x16 pixels). Each of

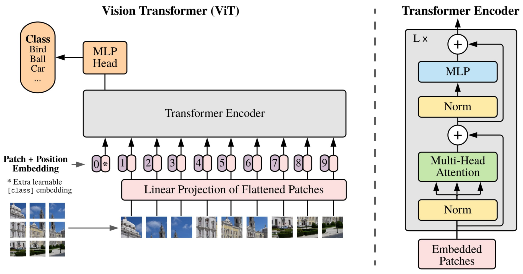

Figure 3: Vision Transformer

these patches is then linearly embedded (flattened and transformed) into a one-dimensional vector of a specified dimension (D).

2. **Positional Embedding:** The order of these patches contains valuable spatial information, which is preserved by adding positional embeddings to the patch embeddings. This results in a sequence of D-dimensional input embeddings, where each input embedding represents the appearance and the relative position of a patch in the image.

3. **Transformer Encoder:** This sequence of input embeddings is then passed through a transformer encoder, which is a stack of identical layers, each containing two sub-layers: a multi-head self-attention mechanism and a position-wise fully connected feed-forward network. A residual connection is employed around each of the two sub-layers, followed by layer normalization.

   - The self-attention mechanism enables each patch to consider all other patches when encoding its representation, allowing for modeling of long-range dependencies.
   - The fully connected feed-forward network, consisting of two linear transformations with a ReLU activation in between, is applied to each position separately and identically.

4. **Classification Head:** After the sequence has passed through the transformer encoder, the transformed embedding of the first position (often called the class token and is initialized randomly and learned during backpropagation) is used to output the classification prediction through a linear layer.

5. **Additional Components:** Some versions of ViT also employ an extra multi-layer perceptron (MLP) head for regression tasks, and stochastic depth regularization for better training efficiency and regularization.

The Vision Transformer design emphasizes the power of transformers to learn from global representations in the data, moving away from the locality principle of Convolutional Neural Networks (CNNs). It also exhibits the flexibility and capability of transformer models to scale with increased data and computational resources, further reinforcing its effectiveness for complex visual tasks.

3.3 Early Fusion: Text and Image Classification

Early fusion approaches attempt to integrate different modalities of a piece of content prior to the classification process, aiming to leverage the interaction between these modalities to enhance the classification performance. In the context of hateful meme classification, this typically involves fusing information from text and image modalities.

Before describing the fusion process, let's discuss the standalone architectures of BERT and ResNet-101[7] that we employ for text and image classification, respectively.

- **BERT:** Bidirectional Encoder Representations from Transformers (BERT) , is a transformer-based machine learning technique for natural language processing. It takes as input a sequence of tokens, and outputs a sequence of contextual token embeddings. The output from the [CLS] token or a pooling of all token embeddings is used as a representation of the entire input sequence.

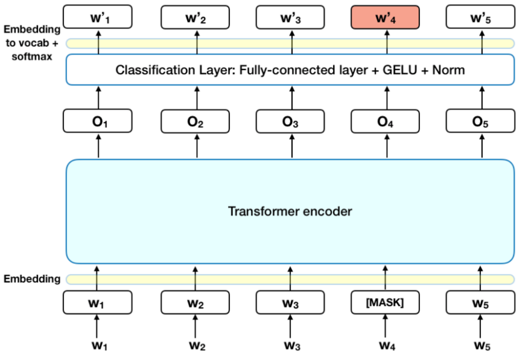
w'₁
w'₂
w'₃
w'₄
w'₅
Embedding to vocab + softmax
Classification Layer: Fully-connected layer + GELU + Norm
O₁
O₂
O₃
O₄
O₅
Transformer encoder
Embedding
w₁
w₂
w₃
[MASK]
w₅
w₁
w₂
w₃
w₄
w₅

Figure 4: BERT

- **ResNet101:** ResNet101 is a variant of the Residual Network (ResNet), which is designed to effectively address the vanishing gradient problem in deep neural networks. The key innovation in ResNet is the introduction of "skip connections" or "shortcuts" that allow the gradient to be directly backpropagated to earlier layers. ResNet101 has 101 layers, and it is widely used in tasks that require the extraction of intricate patterns from images.

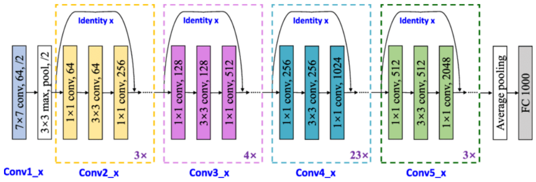

Figure 5: ResNet101

For early fusion, two distinct strategies are proposed:

1. **BERT and Vision Transformer (ViT):** In this strategy, BERT is used for processing the textual data while ViT is used for the image data. The outputs from both models (the [CLS] token from BERT and the class token from ViT) are concatenated together to form a joint representation that contains information from both modalities. This joint representation is then used for classification.

2. **BERT and ResNet101:** The second strategy replaces ViT with ResNet101 for processing the image data. The last layer output of ResNet101, a feature vector, is concatenated with the [CLS] token output from BERT, forming a combined representation for classification. ResNet101 is efficient in extracting hierarchical visual features from images, making it a viable alternative to ViT.

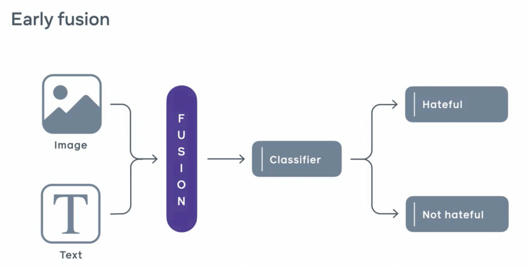

Figure 6: Early Fusion

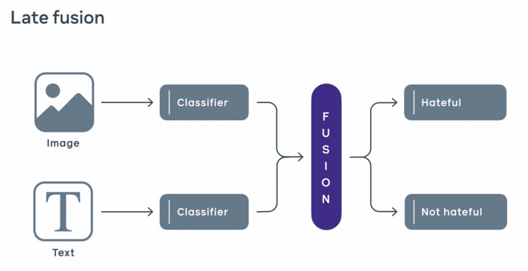

Figure 7: Late Fusion

Both strategies enable the model to consider text and image data concurrently early in the process, thereby improving its ability to detect hateful content that might be difficult to identify by considering each modality separately.

3.4 Late Fusion: Text and Image Classification

Late fusion refers to a strategy where each modality is first classified independently, and then these individual results are combined to yield a final classification. This approach provides more flexibility as each modality can be processed by its most suitable model, though it may not fully capture the complex interplay between the modalities.

One key approaches of late fusion is considered in the context of hateful meme classification:

1. **Sentence Transformer and Vision Transformer (ViT):** This strategy employs the Sentence Transformer for textual data processing and ViT for image processing. Each of these models independently generates a classification score. These scores are then combined (through methods such as averaging, taking a weighted sum, or training a separate classifier on top of the scores) to produce a final decision.

These late fusion strategies provide an effective means of integrating multimodal data for classification. Despite the potential drawback of not fully capturing the interplay between text and images, they leverage the strengths of specialized models for each modality and offer flexibility in combining their results.

4 Experiments

4.1 Dataset

We utilized the Hateful Memes Challenge Dataset provided by Meta, which consists of 8500 training images, a development set of 500 images, anda test set of 1000 images. Each image in the dataset is annotated with a binary label indicating whether the meme is hateful or not. The images in this dataset come from various genres, including cartoons, movie scenes, and internet memes, and the text superimposed on the images ranges from a single word to multiple sentences. In addition, this dataset is significantly imbalanced, with the non-hateful memes significantly outnumbering the hateful ones.

- **Baseline Models:**
  - **Sentence Transformer:** The Sentence Transformer serves as an effective model for text classification tasks. It leverages transformer architectures for the generation of meaningful sentence embeddings. However, its accuracy is relatively low in this task. Given the same text, it's possible that the meme might be hateful or not hateful. Just having the text is not sufficient to make the classification. This is because we have benign confounders where given the same text, the meme can turn from non-hateful to hateful if we change the background image of the meme.
  - **Vision Transformer (ViT):** The Vision Transformer (ViT) introduced a paradigm shift in the computer vision domain, borrowing heavily from the transformers primarily used in Natural Language Processing (NLP). ViT is performing better because even

Table 1: Final model results

Type	Model	Validation		Test	
		Acc.	AUROC	Acc.	AUROC
	Human	-	-	84.70	82.65
Unimodal	Sentence Transformer (ST)	54.80	54.80	55.00	54.60
	Vision Transformer (ViT)	61.60	51.28	52.00	51.10
Multimodal (Unimodal Pretraining)	Early Fusion: BERT + ViT	52.00	52.6	52.6	51.70
	Early Fusion: BERT + ResNet	53.40	53.39	53.20	52.80
	Late Fusion: ST + ViT	52.00	52.00	52.60	52.07

though it's only an image classification model, the text is present within the image itself. The reason that we still see lower accuracy is that because the training dataset is just of 85000 images in the train dataset. ViT, being a sufficiently large model, is able to get overfitted on this dataset. It's expected to perform better if we have a larger dataset.

- **Early Fusion:**
  - **BERT + ViT:** In this strategy, BERT is used for processing the textual data while ViT is used for the image data. The outputs from both models (the [CLS] token from BERT and the class token from ViT) are concatenated together to form a joint representation that contains information from both modalities. This joint representation is then used for classification. However, the performance of this model is not very high, with an accuracy of 50.6% on the test set. This is because there is no attention mechanism applied between the image and the text embeddings as it's done in models like VisualBERT which is expected to improve its accuracy.
  - **BERT + ResNet101:** The second strategy replaces ViT with ResNet101 for processing the image data. The last layer output of ResNet101, a feature vector, is concatenated with the [CLS] token output from BERT, forming a combined representation for classification. ResNet101 is efficient in extracting hierarchical visual features from images, making it a viable alternative to ViT. This model performs slightly better than the BERT + ViT model, with an accuracy of 53.2% on the test set. Resnet performs better than ViT because it's a relatively smaller and the chances of it getting overfit is less compared to ViT.
- **Late Fusion:** Late fusion strategies typically involve training separate models for the image and text data, and then combining their predictions in some way (e.g., by averaging or by training a separate classifier on the predictions). This approach can be more flexible than early fusion, as it allows each modality to be processed independently. However, it may also be less effective at capturing interactions between the modalities. The training time is less for late fusion because the text and the image classification models can be trained parellelly.

5 Conclusion and Future Works:

Classifying hateful memes presents a significant challenge due to the inherent complexity of interpreting nuanced forms of speech such as sarcasm, which are often embedded in these memes. Deep learning algorithms, despite their advanced capabilities, still struggle to accurately detect and interpret these subtleties.

In our future work, we aim to refine our approach by implementing an early fusion strategy with a specific modification for handling images. Given that the text is already separately available in our dataset, we plan to remove the captions from the images. This step is intended to eliminate redundancy in the data, which we anticipate will enhance the performance of our model.

By reducing this duplication, we expect to see improvements in our test accuracies and Area Under the Receiver Operating Characteristic (AUROC) values, thereby increasing the effectiveness of our hateful meme classification efforts.

References

[1] Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10, 2017.

[2] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515, 2017.

[3] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, page 1304113049, 2020.

[4] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.