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 a critical problem in the field of computer vision that aims to automatically identify and categorize memes containing offensive or harmful content. With the rapid growth of social media platforms, memes have become a popular medium for conveying ideas, humor, and opinions. Traditional text-based analysis methods have not been able to fully capture the semantic relationship between the text and image, while image-based techniques might miss the context provided by the text. To tackle this problem, we will be exploring approaches that combine textual and visual features, utilizing deep learning techniques such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), and the more recent Transformer-based models like BERT and ViT.

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:** A comprehensive definition of hateful memes is vital, considering various forms, context, and cultural factors influencing their interpretation.
- **Multimodal Feature Engineering:** Developing techniques to capture interplay between textual and visual features, as well as their relationships, is critical to differentiate hateful and non-hateful memes.
- **Dataset Creation and Annotation:** Curating a diverse and representative meme dataset is essential for training and evaluating classification models, with attention to defining hateful memes and potential annotation biases.
- **Model Evaluation and Performance Metrics:** Selecting suitable metrics for evaluating classification models is crucial, accounting for specific challenges like dataset imbalance and false positives/negatives while identifying 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.

2 Methods

In this section, we will be discussing some of the methods that we will be implementing as part of our project.

- **Data Preprocessing:** The collected data will be processed to remove any noise. The text will be extracted from the image using OCR recognition. This is inspired from [4] .
- **Late fusion:** We will be training the late fusion architecture from scratch using Convolutional Neural Network for extracting the image embedding and LSTM for extracting the text embedding. We will concatenate the text and the image embedding, passing it through a fully connected layer and the output will be whether the meme is hateful or not.
- **Transfer Learning:** We will be using existing pretrained transformer[5]-based models such as BERT[1], which is trained on large scale text corpora for extracting the text embeddings and ViT[2] for extracting the image embeddings. We will fine-tune the model to classify hateful memes.
- **Evaluation** We will be using standard evaluation metrics such as precision, recall and F1 score. We will not be using accuracy since the dataset is imbalanced with majority of the memes being non hateful.

3 Experiments

We will be using Hateful Memes Challenge Dataset[1] provided by Meta. It is an open source dataset designed to measure progress in multimodal vision and language classification. It consists of 8500 train images, a dev set of 500 images and a test set of 1000 images. The dataset also provides with the image caption, the path to the image and label (0 = non hateful and 1 = hateful) in train.jsonl and test.jsonl for training and testing respectively. The dataset is challenging because it requires understanding of both the visual and textual aspects of the memes and some of the memes are ambiguous or sarcastic. Also, some of the memes change from being non hateful to hateful upon changing the context. They are called benign cofounders.

We will be implementing the architectures mentioned above to improve the performance on hateful meme classification. We will also be performing ablation studies to explore the contribution of each part of the architecture towards performance such as removing the image embeddings and just using the text embedding to see how much there is a drop in F1 score. The work in [3] shows that language ends up doing most of the heavy lifting in multimodal reasoning. This ends up in baseline language models performing quite well and image not contributing much. The dataset consists of tricky memes where information needs to be extracted from both image and text to classify hateful memes.

References

[1] 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.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[3] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. 2020.

[4] Niklas Muennighoff. Vilio: State-of-the-art visio-linguistic models applied to hateful memes. *arXiv preprint arXiv:2012.07788*, 2020.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[1] https://ai.facebook.com/tools/hatefulmemes/