Multimodal Hate Speech Detection

15.095 Machine Learning Under a Modern Optimization Lens

To what extent does the incorporation of both image and text data enhance the accuracy of hate speech classification in tweets compared to using either modality alone?

Submitted by: Pavena Vongkhammi, Raghav Raahul Manoharan Jayanthi

1. Context and Challenge: An Overview

In the rapidly evolving landscape of social media, platforms such as X (formerly Twitter) have emerged as powerful mediums for global expression, allowing individuals a platform to share thoughts, ideas, and experiences. However, this unprecedented connectivity has brought to light a concerning challenge - the amplification of hateful speech and offensive content. Hate speech can be defined as any communication using offensive language to target individuals or groups based on identity factors such as religion, ethnicity, nationality, race, color, gender, or other characteristics, poses a significant threat to online safety and well-being.

The prevalence of hate speech on social media platforms can have lots of detrimental effects. It contributes significantly to the creation of division, fear, and societal tension, fostering discrimination and marginalization. Beyond its societal impact, hate speech inflicts psychological harm, causing emotional distress and fear among targeted individuals and communities. Moreover, the surge in hate speech raises profound legal and ethical questions, touching upon issues of free speech, censorship, and discrimination.

In response to this critical issue, our project aims to develop a machine learning model that can effectively differentiate between hateful and non-hateful tweets. This model will leverage both textual and image data associated with each tweet, employing a multimodal approach to comprehensively analyze and classify content. By utilizing the power of artificial intelligence in conjunction with both text and image features, our objective is to provide a robust solution for combating the spread of hate speech on social media platforms.

The significance of this project extends beyond mere content moderation. We envision a tangible improvement in online safety and user experience, allowing individuals to curate their social media feeds to be free from hateful content. By offering users the option to filter out such content, we aim to create a safer, more inclusive online environment. Additionally, the successful implementation of this model has the potential to enhance the reputation of X as a responsible and proactive company in the fight against hate speech.

As we delve into the development of this machine learning model, we acknowledge the complex nature of the problem and the multifaceted impact of hate speech. Our commitment is not only to the technological advancement of content moderation but also to the ethical considerations inherent in navigating the delicate balance between free speech and the prevention of harm. Through this initiative, we aspire to contribute to a digital landscape where individuals can express themselves freely, without fear of discrimination or psychological harm.

2. Dataset

The MMHS150K Dataset, available on Kaggle, serves as a meticulously curated and annotated collection that uniquely combines text and images in the context of hate speech. This dataset, consisting of 150,000 tweets, has been systematically gathered to facilitate the study of multimodal hate speech, where each entry is annotated to provide valuable insights into the intricate relationship between textual and visual content.

The dataset is divided into three sets: a training set with approximately 135,000 samples, a testing set comprising 10,000 samples, and a validation set containing 5,000 samples. The meticulous curation ensures that the dataset is well-balanced and representative of various instances of hate speech discussions on social media.

The data collection process utilized the Twitter API to acquire real-time tweets spanning the period from September 2018 to February 2019. Tweets were selected based on their inclusion of any of the 51 identified Hatebase terms, which are known to be prevalent in hate speech discussions, as established in prior research. Rigorous filters were applied to ensure the quality of the dataset, excluding retweets, tweets with fewer than three words, and those containing explicit content.

The annotation process was carried out using Amazon Mechanical Turk, a crowdsourcing platform. Annotators were provided with a clear definition of hate speech, accompanied by illustrative examples to ensure a shared understanding of the task. Each tweet, along with its associated image, was classified into one of six categories: no attacks, racist, sexist, homophobic, religion-based attacks, or attacks targeting other communities. To mitigate potential discrepancies, each tweet underwent assessment by three distinct annotators.

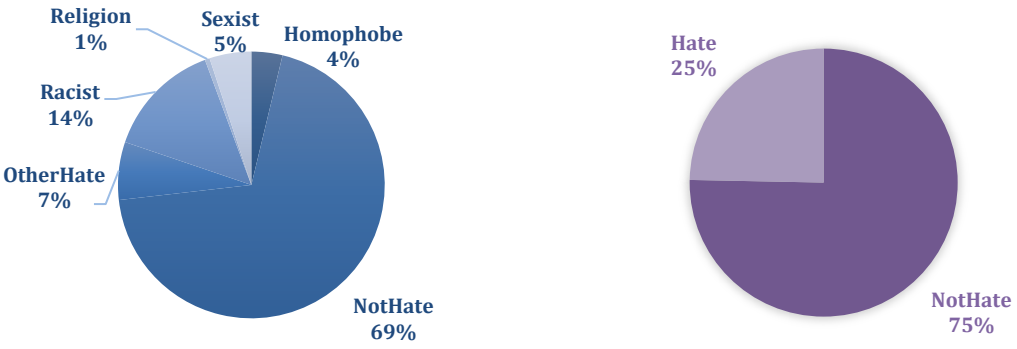

Figure 1: Label Distribution

We receive three distinct assessments for every tweet from different reviewers, and we determine the ultimate result as a binary classification (Hate or NotHate) by following a

majority voting process (at least 2 out of the 3 reviewers must label the tweet as NotHate). For each tweet, we check if NotHate is the majority label. If this is the case, we classify the id (as well as its correspnding image and text) as notHate. Else, the id is classified as Hate. For instance, if a tweet is categorized as [Religion, Racist, NotHate], the final outcome will be classified as Hate since NotHate is not the majority element; conversely, if the tweet's classification is [NotHate, Racist, NotHate], the final outcome will be deemed NotHate. Approximately 25% of the tweets are categorized as Hate speech, while 75% are categorized as NotHate speech.

Dataset Content

		
big d*** energy	Women always want to set the bar high.... Can't you cheat with n***** on my level?	Arab Spring's Legacy: Islamist Gang Terror

Figure 2: Dataset Content

3. Methodology

In this study, we employ two distinct methodologies to develop a robust multimodal machine learning model for the classification of hate speech on social media. The first method utilizes predicted probabilities derived from pre-trained models, while the second method leverages embeddings obtained from the same models. Both approaches address the class imbalance in the training data by downsampling to ensure equal representation of hate and non-hate classes.

Method 1: Predicted Probabilities

1.1 Data Preprocessing:
To mitigate the impact of class imbalance, the training data is downsampled, resulting in 29,477 images in each class (hate and not hate). This step ensures that the model is exposed to a balanced set of examples during training.

1.2 Image and Text Model Fine-Tuning:
Pre-trained CNN models VGG16 and ImageNet are fine-tuned and then trained to classify images into binary classes, generating probability predictions. Similarly, pre-trained text models BERT and word2vec are fine-tuned to classify text into binary classes, producing probability predictions.

1.3 Probability Prediction Combination:
The probability predictions from both the text and image models are combined using four classification models: Logistic Regression, XGBoost, CatBoost, and LightGBM. This ensemble approach aims to capture the complementary information present in both modalities, enhancing the overall predictive power of the model.

1.4 Final Prediction:
The combined probability predictions are used to make a final binary classification, with Class 1 denoting Hate and Class 0 denoting Not Hate. This approach allows for a nuanced understanding of the model's confidence in its predictions.

Method 2: Embeddings

2.1 Data Preprocessing:
Similar to Method 1, downsampling of the training data is performed to address class imbalance, ensuring 29,477 images in each class (hate and not hate).

2.2 Image and Text Model Fine-Tuning:
Pre-trained CNN model ImageNet was finteuned and then trained to generate image embeddings. Simultaneously, pre-trained BERT was used to generate text embeddings.

2.3 Embedding Combination:
The embeddings from both text and image models are combined and fed into four classification models: Logistic Regression, XGBoost, CatBoost, and LightGBM. This approach captures the semantic relationships within and between modalities, allowing for a comprehensive representation of multimodal information.

2.4 Final Prediction:
Similar to Method 1, the combined embeddings are used to make a final binary classification, with Class 1 denoting Hate and Class 0 denoting Not Hate.

Rationale:

- Class Imbalance Mitigation: Downsampling addresses the inherent imbalance in hate and non-hate classes, ensuring fair representation during model training and improving model learning across both outcomes. Imbalanced datasets can result

in models that are biased towards the majority class, making them less effective in identifying and correctly classifying instances from the minority class. Downsampling also helps address overfitting. In situations where the majority class dominates the dataset, machine learning models might focus excessively on learning patterns associated with that class. Downsampling prevents overfitting by providing a more balanced training set, reducing the likelihood that the model will memorize the majority class examples.

- Predicted Probabilities vs. Embeddings: Predicted probabilities provide a probabilistic understanding of the model's confidence in its predictions, while embeddings capture semantic relationships within and between modalities, offering a nuanced representation of multimodal data.
- Ensemble Approach: The ensemble of multiple classification models ensures robustness and mitigates the risk of model bias by leveraging the strengths of individual algorithms.

By employing these two distinct methodologies, we aim to develop a comprehensive multimodal model for hate speech classification, contributing to the improvement of online safety and user experience on social media platforms.

4. Results

We first engaged pre-trained models, subsequently fine-tuning them with our dataset. This fine-tuning process enabled us to extract embeddings for both textual and image data, subsequently passing these embeddings through a neural network, which provided probability predictions as the input for method 1. The subsequent tables display the performance of the fine-tuned models.

	Accuracy	**AUC**
BERT	67.56%	71.97%
Word2Vec	67.08%	71.48%

Table 1: Text Performance

	Accuracy	**AUC**
VGG16	49.99%	50.00%
ImageNet	50.02%	50.00%

Table 2: Image Performance

Method 1: Predicted Probabilities

In our primary approach, to synthesize a comprehensive outcome, we merged the probabilities derived from both text and image models and utilized them as inputs into the classification models to yield the final result. ImageNet and BERT were chosen for combined predicted probabilities model because of their superior performance on image and text data alone respectively.

	Accuracy	**AUC**
Logistic Regression	67.56%	71.97%
XGBoost	67.37%	71.98%
CatBoost	67.30%	71.98%
LightGBM	67.42%	71.98%

Table 3: Final Performance (1st Approach)

While text-based analysis demonstrates proficient performance, the analyzing images presents inherent challenges. The standalone prediction of hate solely from images proves intricate; instances labeled as "NotHate" may contain latent cues indicative of hate, potentially obscured due to misclassification, highlighting the need for nuanced approaches in image-based prediction to capture subtle yet crucial visual cues.

Relying solely on text yielded superior results, achieving an accuracy of approximately 67–68%, surpassing the accuracy of using only images by about 36%. Interestingly, combining text and images resulted in a lower performance compared to using text alone. This disparity arises from the image model's inability to capture hate patterns effectively.

Method 2: Embeddings

Our secondary approach diverged by combining the embeddings derived from training the ImageNet and BERT models on the textual and image data respectively. ImageNet and BERT were chosen because of their superior performance on image and text data alone respectively. By fusing these embeddings, we crafted a unified representation that was then directly fed into the classification models, fostering an alternative strategy to analyze and produce the conclusive outcome.

	Accuracy	**AUC**
Logistic Regression	63.48%	67.86%
XGBoost	61.18%	64.97%
CatBoost	62.21%	66.33%
LightGBM	61.45%	65.57%

Table 4: Final Performance (2nd Approach)

The performance of the model in the second approach remains subpar, primarily due to the increased complexity caused by merging word and image embeddings. This complexity poses challenges in capturing discernible patterns, culminating in an overall accuracy of approximately 62% across all classification models.

The model's performance is significantly impacted by another critical factor: the subjective nature of the labeling strategy employed by reviewers. This subjectivity adds complexity, hindering the model's capacity to accurately discern authentic patterns within the data.

5. Discussion

Overall, from this project, we've observed that the result of combining images and text to generate class predictions has a lower accuracy and AUC compared to text alone which deviates from our anticipated outcome of better outcomes with multimodal data and there could be a few reasons to explain this discrepancy:

Reevaluation of Image Classification:

The unexpected results obtained may prompt a reexamination of our image classification strategy. One key insight is the subjectivity inherent in the labels assigned to images. To address this, we propose refining the image classification approach by considering an image as "NotHate" only if all assigned labels unanimously indicate "NotHate." This adjustment aims to mitigate potential discrepancies in subjective labeling and provide a more objective basis for image classification.

Challenges with Subjective Labeling:

The subjective nature of human annotators introduces variability in labeling, particularly in the context of hate speech classification. The interpretation of hate speech can vary among individuals, leading to discrepancies in the assigned labels. Given the intricacies of hate speech, which often involve subtle nuances, subjective labeling can contribute to inconsistencies in the training data.

6. Potential next steps:

Unanimous NotHate Labels for Image Classification:

To enhance the objectivity of image classification, we propose considering an image as "NotHate" only if all assigned labels unanimously indicate "NotHate." This approach

aligns with the notion that images should be classified as non-hateful only when there is unanimous agreement among annotators.

Exploring Objective Image Classification:

A more objective image classification approach may enhance the model's ability to discern non-hateful content. Perhaps, coming up with a set of defined rules to enforce objectivity can be a good starting point. Objective classification tasks, such as distinguishing between cats and dogs, showcase the effectiveness of images in providing clear and unambiguous information. Hate speech, however, involves complex linguistic and contextual nuances that can make subjective labeling challenging.

Individual Contributions:

Raghav

- Finetuned and ran pretrained CNN and VGG16 models on the image data for classification with all attempts to improve accuracy such as image preprocessing, standardization, noise reduction, experimenting with different number of layers, activation functions, number of units in each layer, dropout layers, epoch count, step size etc (method 1); got prediction probabilities
- Got image embeddings (method 2)
- Ran logistic regression and xg boost models combining prediction probabilities from images and text (model 1) and combining embeddings from images and text (model 2)
- Wrote introduction, problem statement, methodology, and discussion sections of the report
- Did methodology, intro, and problem statement of powerpoint slides.
- Found the dataset

Hannah

- Explored various datasets but ultimately utilized this dataset called MMHS150K from Kaggle for our analysis
- Conducted data preprocessing, including the removal of URLs, account names, etc. Implemented downsampling to prevent overfitting and shared the IDs with Raghav for consistent use in the Image model
- Responsible for the text-based analysis, fine-tuning the BERT model for word embedding extraction. Additionally, employed Word2Vec for further word embedding generation
- Developed two distinct methodologies:
    - Constructed a model utilizing text-based probabilities, followed by employing classification models (CatBoost and LightGBM) for final predictions
    - Built classification models (CatBoost and LightGBM) using combined embeddings from both text and image sources to get final predictions

- Delivered a comprehensive presentation encompassing introduction, data overview, methodology, key findings, significant insights, and proposed next steps
- Authored a detailed report covering dataset specifics, results, and comprehensive discussions around the findings