Multimodal Hate Speech Detection!

Team Members:
Pavena Vongkhammi
Raghav Jayanthi

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****?*

What is hate speech? Why do we care?

*Any kind of communication in speech, writing or behavior that uses **offensive language** targeting a person or group based on their identity (religion, ethnicity, nationality, race, color, gender or other identity factor).”*

Impact of Hate Speech


Social

Create division, fear, and societal tension by fostering discrimination and marginalization.


Psychological

Cause emotional distress, fear, and psychological harm to targeted individuals or communities


Legal and Ethical

Raises legal, ethical questions on free speech, censorship, discrimination

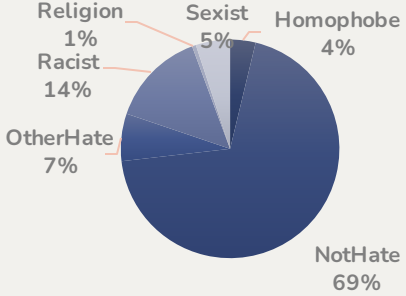
Dataset

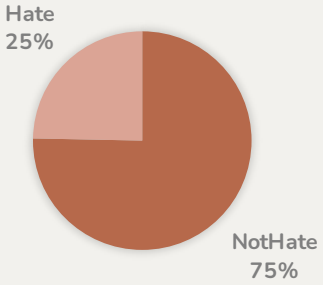
The MMHS150K Dataset

150,000 tweets (containing text and image) from September 2018 to February 2019

Train	Validation	Test
135,000 samples (90%)	5,000 samples (3%)	10,000 samples (7%)

Distribution





Example

big d*** energy

Homophobe
Homophobe
Racist

Arab Spring's Legacy: Islamist Gang Terror

Racist
Racist
Religion

Methodology

Method 1: Predicted probabilities

Classify tweets into binary categories (hate and not hate) using majority vote from three reviewers

Down sampled trai ning data

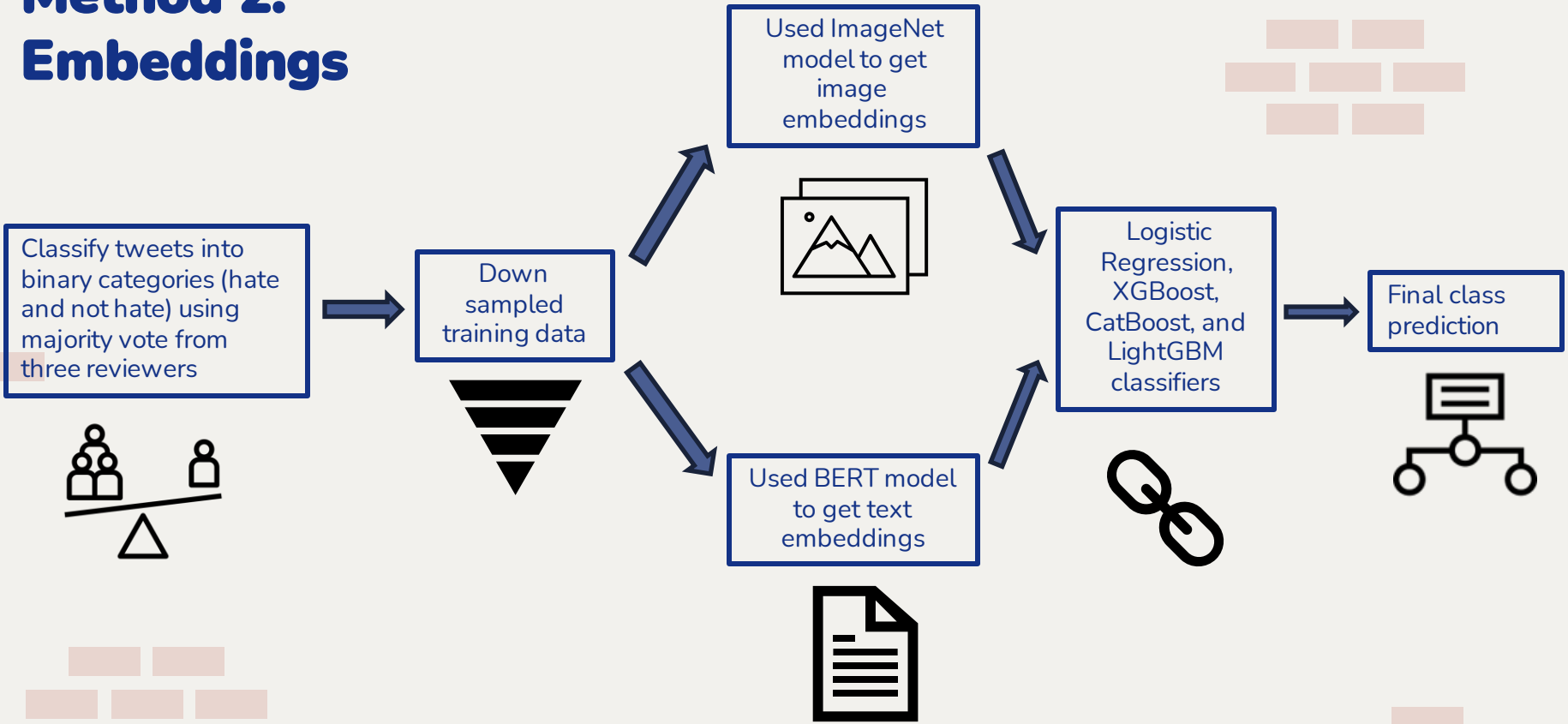
Finetuned VGG16 and ImageNet to classify images and get probability predictions

Finetuned BERT and word2vec to classify text and get probability predictions

Logistic Regression, XGBoost, CatBoost, and LightGBM classifiers

Final class prediction

Method 2: Embeddings



Key Results

Singular modality predictions

Text

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

Image

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

Method 1: Predicted Probabilities

Final Prediction (Text + Image)

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

Method 2: Embeddings

Final Prediction (Text + Image)

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

Which is the preferred method?

Method 1 (using predicted probabilities)

Which is the preferred classification method for combining text and images?

Logistic regression

Is text + images superior to either text or either images?

Text + image is just as good as using only text indicating images were not helpful in this application

Key Insights & Next Steps

Key Insights

Performance

Text analysis excels, but images are tricky. Predicting hate from images alone is tough - some labeled 'NotHate' might hide hate cues, leading to misclassification

Thanks "— r*t****" For Following Me !!!

Label: [Sexist, NotHate, NotHate]

Subjective Label

Subjective labeling complicates the model. Some images marked as 'NotHate' might be considered 'Hate' by others, complicating accurate classification

HAHA!

Next Steps!

to enhance model's performance

Exploring alternative labeling criteria:

Instead of majority voting, we can consider labeling images as non-hateful only when there's unanimous agreement from all three reviewers

Exploring Objective Image Labeling Methods:

An objective approach improves identifying non-hateful content. Defining clear rules could kickstart this shift

Thanks

Does anyone have any questions?