Hateful Memes Classification

Team 6

Vaibhav Garg

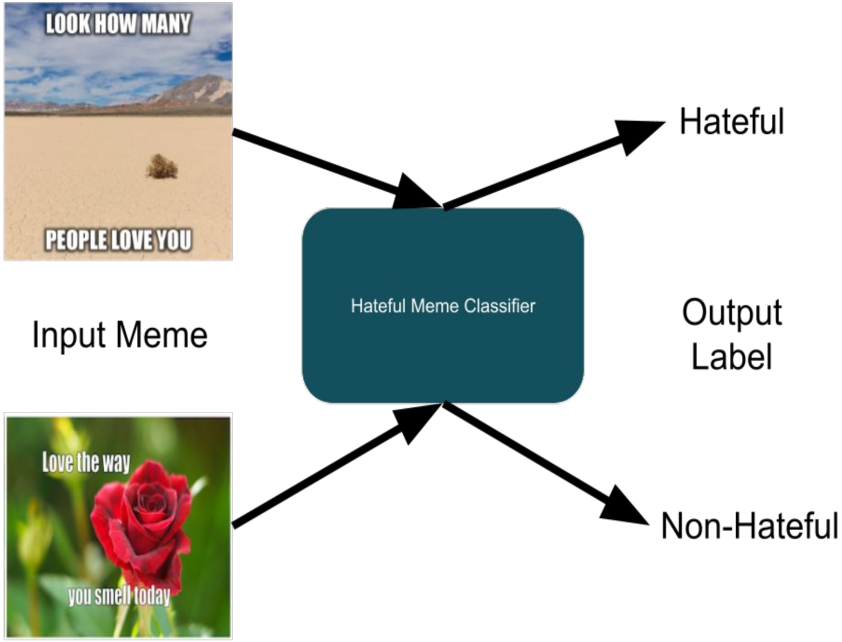
Sagar Joshi

Preet Thakkar

Akshay Goindani

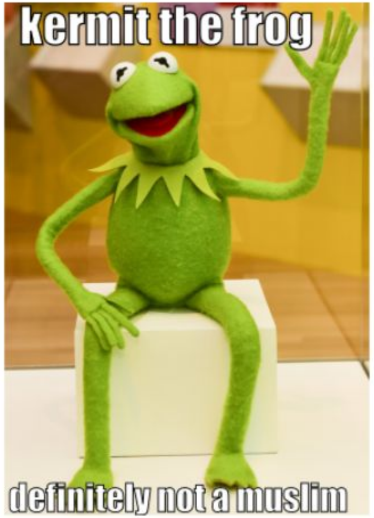
Problem Statement

The goal is to predict whether a meme is hateful or non-hateful. This is essentially a binary classification problem with multimodal input data consisting of the the meme image itself (the image mode) and a string representing the text in the meme image (the text mode).



Need?

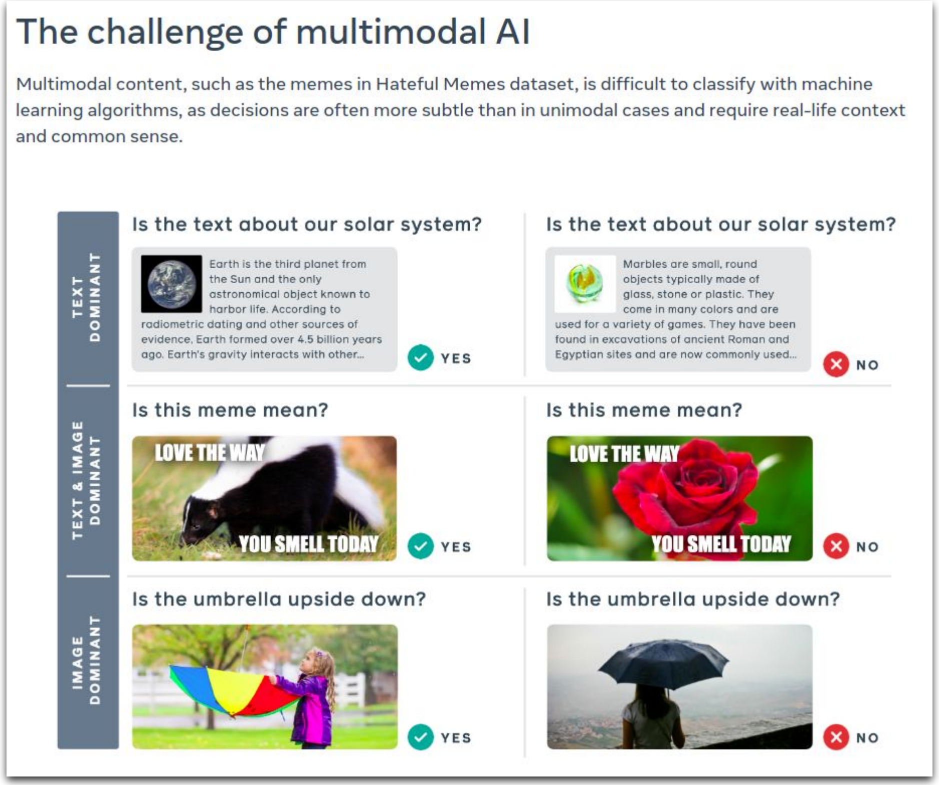
Internet memes have become a major form of communication and expression, they're an essential part of social media's popular culture. But like with any other type of content in social media, they also can lead to the spread of hate speech and propaganda.



Before and after you came into my life

The Facebook AI Challenge

Facebook AI's Hateful Memes Challenge was one of the very first large-scale attempts to explore this problem and the results prove just how challenging this problem is, interpreting a meme requires a lot of cultural, political, and contextual knowledge and multimodal understanding since it is both the text and the image combined that brings out the real meaning of the meme.



The challenge of multimodal AI

Multimodal content, such as the memes in Hateful Memes dataset, is difficult to classify with machine learning algorithms, as decisions are often more subtle than in unimodal cases and require real-life context and common sense.

Is the text about our solar system?

Earth is the third planet from the Sun and the only astronomical object known to harbor life. According to radiometric dating and other sources of evidence, Earth formed over 4.5 billion years ago. Earth's gravity interacts with other...

Is the text about our solar system?

Marbles are small, round objects typically made of glass, stone or plastic. They come in many colors and are used for a variety of games. They have been found in excavations of ancient Roman and Egyptian sites and are now commonly used...

Is this meme mean?

Is this meme mean?

Is the umbrella upside down?

Is the umbrella upside down?

Dataset Provided by Facebook AI

- Facebook AI released a novel Hateful Memes Dataset consisting of 10K memes
- It is carefully constructed with the help of human annotators so that unimodal architectures will fail
- Presence of benign confounders

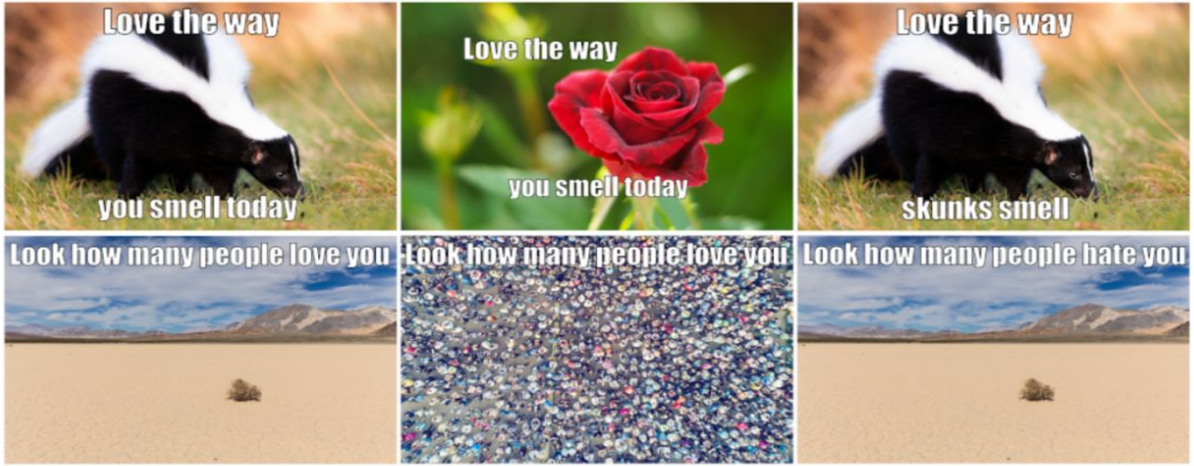

Love the way
you smell today
Love the way
you smell today
Love the way
skunks smell
Look how many people love you
Look how many people love you
Look how many people hate you

Fig: The original hateful memes are in the left column. Image confounders formed by replacing the image are in the middle column. Text confounders formed by minor replacements to the original text are in the right column.

Solutions

There are multiple solutions for designing the Hateful Meme Classifier. Broadly, these solutions are classified into two classes -

1. **Unimodal** Solutions : In these type of solutions, only one modality (either Image or Text) is used to predict the label.
2. **Multimodal** Solutions : In these type of solutions, both the modalities are used to predict the label.

In both unimodal and multimodal approaches, the inputs to the classifier are image and text features. The image and text features are extracted using various feature extraction methods. Based on the design of the classifier, the image feature extraction method varies whereas the text feature extraction is more or less the same for all classifiers.

Why Multimodal Approach?

Memes contain both image and text, and the sentiments from the meme can only be interpreted if information from both image and text are captured together. Hence, unimodal approaches that just focus on one of the modalities (image or text), perform very poorly on hateful meme classification task.

The Facebook's Hateful Meme Dataset also contains benign confounders and these are added because in such a scenario, unimodal models that focus only on the textual or visual modality will fail and only the multimodal models will be able to learn true reasoning.

In our work, we will focus on multimodal approaches. Based on our literature survey, we found two multimodal methods that achieve performance comparable to humans. These methods are described in -

1. Niklas Muennighoff. Vilio: State-of-the-art visio-linguistic models applied to hateful memes.
2. Riza Velioglu and Jewgeni Rose. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge.

TYPE	MODEL	TEST	
		Acc.	AUROC
	Human	84.70	82.65
Unimodal	Image-grid	52.00	52.63
	Image-region	52.13	55.92
	Text BERT	59.20	65.08
Multimodal *(Unimodal pretraining)*	Late fusion	59.66	64.75
	Concat BERT	59.13	65.79
	MMBT-grid	60.06	67.92
	MMBT-region	60.23	70.73
	ViLBERT	62.30	70.45
	Visual BERT	63.20	71.33
Multimodal *(Multimodal pretraining)*	ViLBERT CC	61.10	70.03
	Visual BERT COCO	64.73	71.41

Image Feature Extraction

There are multiple image feature extraction methods available. In case of memes, the image features are not extracted using just the pixel values, but features of different objects in the image are extracted using object detection models. Different methods to extract object / region based features -

1. Extracting region based features using ResNeXT-152 based Mask-RCNN model. This model outputs 2048 Dimensional features for a single region of interest. 100 such regions are extracted.
2. Extracting RoI features using bottom-up attention in detectron2 framework. Different feature sets with varying no of RoIs (36, 50, 72, 100) are created.

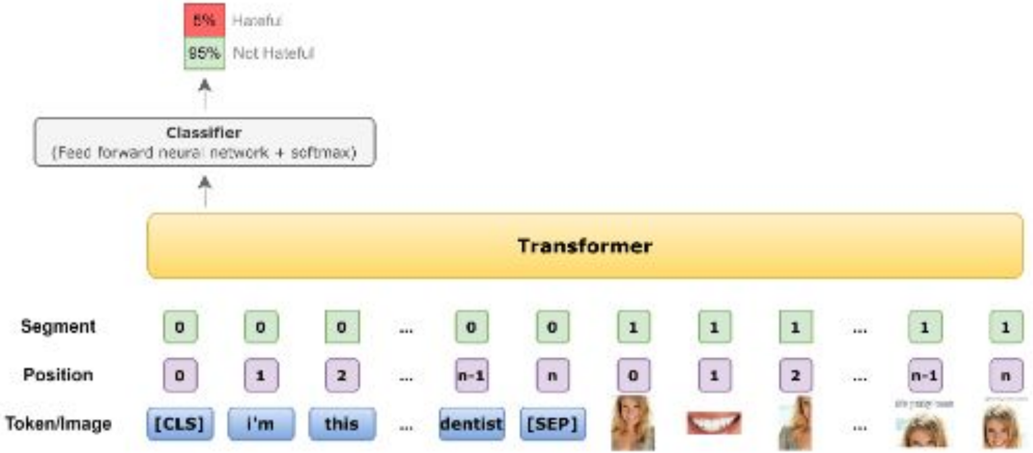
VisualBert

VisualBert is a multimodal BERT for vision and language. Our solution using VisualBert is described as follows

1. **Dataset Expansion** - In addition to the facebook's hateful meme dataset, 328 memes from memotion dataset were added to the training data. These 328 memes were selected based on the similarity with the memes in the facebook's hateful meme data.
2. **Image Encoding** - We used ResNeXT-152 based Mask-RCNN model to get region based image features. Before using these image features, we project them into the textual embedding space, so that image and text are in the same space.
3. **Training** - In this phase, image features and text features (extracted using BERT) are combined using transformer. The self-attention in the transformer architecture allows the model to jointly look at image and text features.
4. **Classification** - The output of the transformer is passed through a classification layer (fully connected layer) to calculate the probability of the input meme being hateful.

Ensemble

1. The VisualBert model described in the previous slide is trained 96 times with different values for different hyperparameters.
2. Out of these 96 models, best 27 models are extracted based on the accuracy on the development set.
3. An ensemble of these 27 models is created. Majority voting is used to make the final prediction.

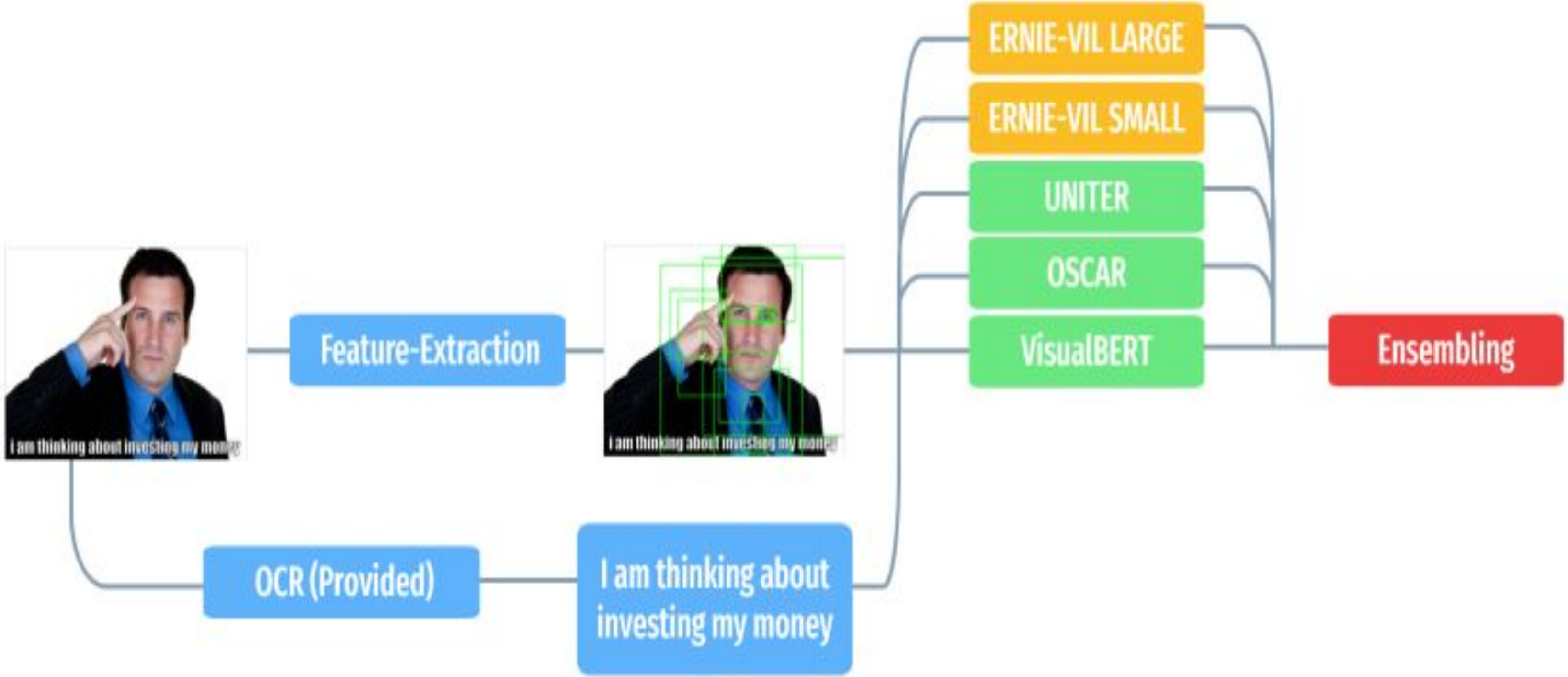


Results for VisualBert Ensemble

Dataset	Accuracy	AUROC
Dev Seen	71.2 %	0.81
Dev Unseen	71.3 %	0.75
Test Seen	66.2 %	0.74

Vilio Model

- With the presence of benign confounders, it is evident that any approach will be using Vision+Language machine learning models (which, are based on transformer architecture).
- Types of Vision+Language machine learning models:
  - Single-stream architecture: These rely on a single transformer pipeline to process the image and language input at the same time.
  - Dual-stream architecture: They use separate transformer pipelines for vision and language, in which the the two modalities interact through cross-attention.
- In this solution, a combination of single-stream and dual-stream architectures is used.



Feature-Extraction

OCR (Provided)

I am thinking about investing my money

ERNIE-VIL LARGE

ERNIE-VIL SMALL

UNITER

OSCAR

VisualBERT

Ensembling

i am thinking about investing my money
i am thinking about investing my money

Approach

- Image Encoding:
  - The models used Regions of Interest extracted using bottom-up attention model based on training of Faster R-CNN with ResNet-101, using object and attribute annotations from Visual Genome.
  - The above implementation available in detectron2 framework is used.
  - Different feature sets with varying no of RoIs (36, 50, 72, 100) are created.
  - Together with the meme text, which has been extracted using optical character recognition (OCR) and provided in the dataset, features are then fed into the models.
- Training:
  - Pre-trained weights provided by the original authors of the V+L models are used.
  - For each model, different strategies were used for pre-training phase, the details of which are provided in the paper.
  - All models are then finetuned using binary cross-entropy loss, Adam optimizer, .
  - The weights from the last step are taken for all models and used for inference on the test set.
  - The results are processed in an ensembling loop applying simple averaging, rank averaging, power averaging and simplex optimization to produce the final predictions.
  - Overall, not much time was spent on hyperparameter optimization, as fundamental architecture changes have a larger impact.

Results

Model	AUROC
VisualBERT	0.75
OSCAR	0.76
UNITER	0.77
ERNIE-ViL Small	0.79
ERNIE-Vil Large	0.81

Individual Model Performance

Dataset	AUROC	Accuracy
Dev Seen	0.81	73.2%
Dev Unseen	0.80	72.0%
Test Seen	0.83	73.6%

Ensemble Performance

Additional Models

We tried adding two new models to the ensemble to check for performance improvement: DeVLBERT & LXMERT

Both the architectures are dual-stream multimodal transformers.

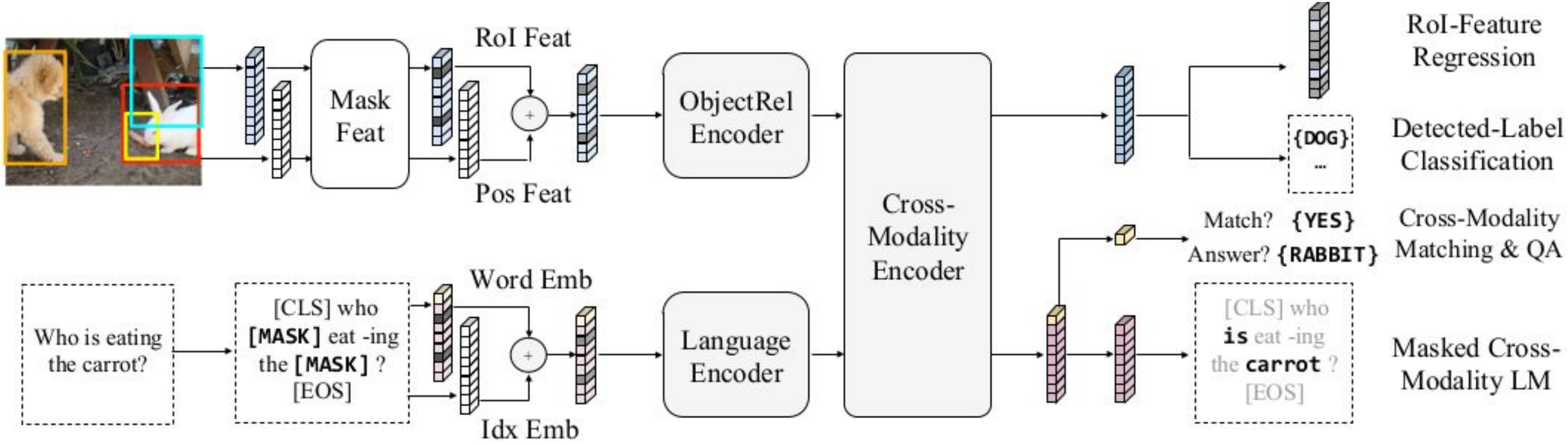
Model	AUROC
DeVLBERT	0.7744
LXMERT	0.7442

DeVLBERT: Deconfounded Visio-Linguistic BERT

The architecture proposes to mitigate the issue of spurious correlations and improving the generalization ability of the model when pretrained on out-of-domain data distribution.

The model is pretrained on Conceptual Captions dataset, similar to ViLBERT using MLM and MOM objectives.

LXMERT: Learning Cross-Modality Encoder Representations from Transformers



The model is pretrained using data from 5 different datasets on 5 different pretraining tasks: Masked Cross-Modality LM, RoI-Feature Regression, Detected-Label Classification, Cross-Modality Matching, Image Question Answering

Results

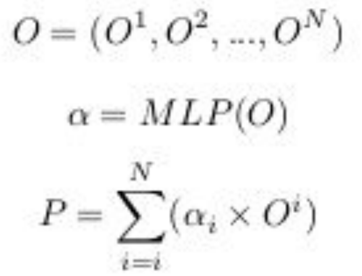
Ensemble	Dataset	AUROC
VisualBERT, UNITER, OSCAR, ERNIE-ViL	Test Seen	0.8254
VisualBERT, UNITER, OSCAR, ERNIE-ViL, LXMERT, DeVLBERT	Test Seen	0.8221

Why did Vilio perform so much better?

- Use of better feature extraction for images
    - Use of bottom-up attention in detectron-2 framework
- Diversity in feature set per architecture
    - Use of 3-5 different feature sets for training models for each architecture
    - Simple-averaged predictions of the trained models are used for further ensembling
- Diverse model architectures
    - Single-stream architectures: OSCAR, UNITER, VisualBERT
    - Dual-stream architectures: ERNIE-Vil - Small & Large - using models pretrained on CC / VCR datasets
- Task-adaptive pretraining
    - OSCAR was pretrained on Image-text matching, MLM
    - VisualBERT was pretrained on MLM
- Better training strategies and ensembling
    - Stochastic Weighted Averaging for last 25% of training
    - Simplex Optimization for final ensembling

Mixture of Experts

1. Mixture of Experts (MOE) is an ensemble technique. In MOE, N different models are combined which are also known as experts.
2. The final probability of a meme being hateful is a weighted sum of the output of the individual models. The weights in the weighted sum are calculated using a Multi Layer Perceptron (MLP).
3. The final probability is used to calculate the loss and the loss is back propagated to train the MLP.
4. Here, O is the output probabilities from all the models and P is the final probability.



$$O = (O^1, O^2, ..., O^N)$$

$$\alpha = MLP(O)$$

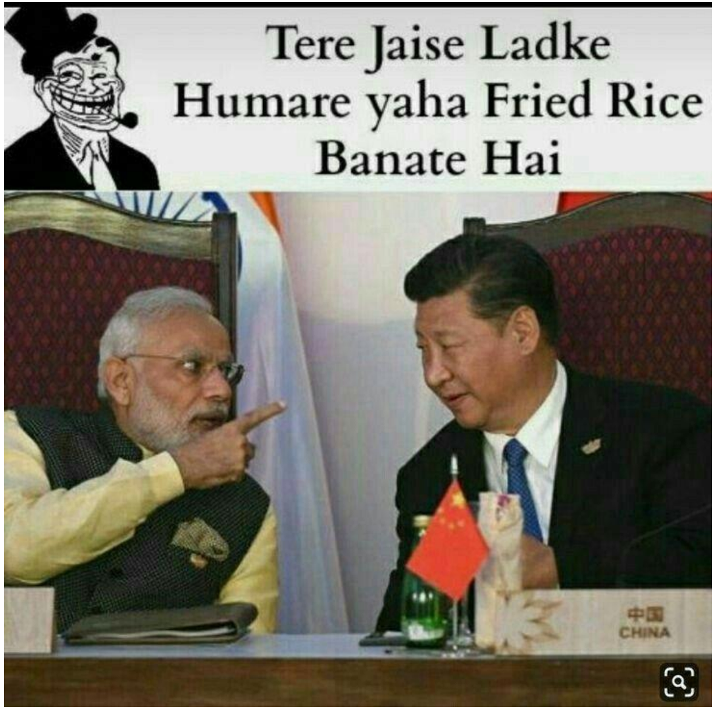
$$P = \sum_{i=i}^{N} (\alpha_i \times O^i)$$

Results using MOE Ensemble

Dataset	Accuracy	AUROC
Dev Seen	73.6 %	-
Test Seen	79.5 %	0.732

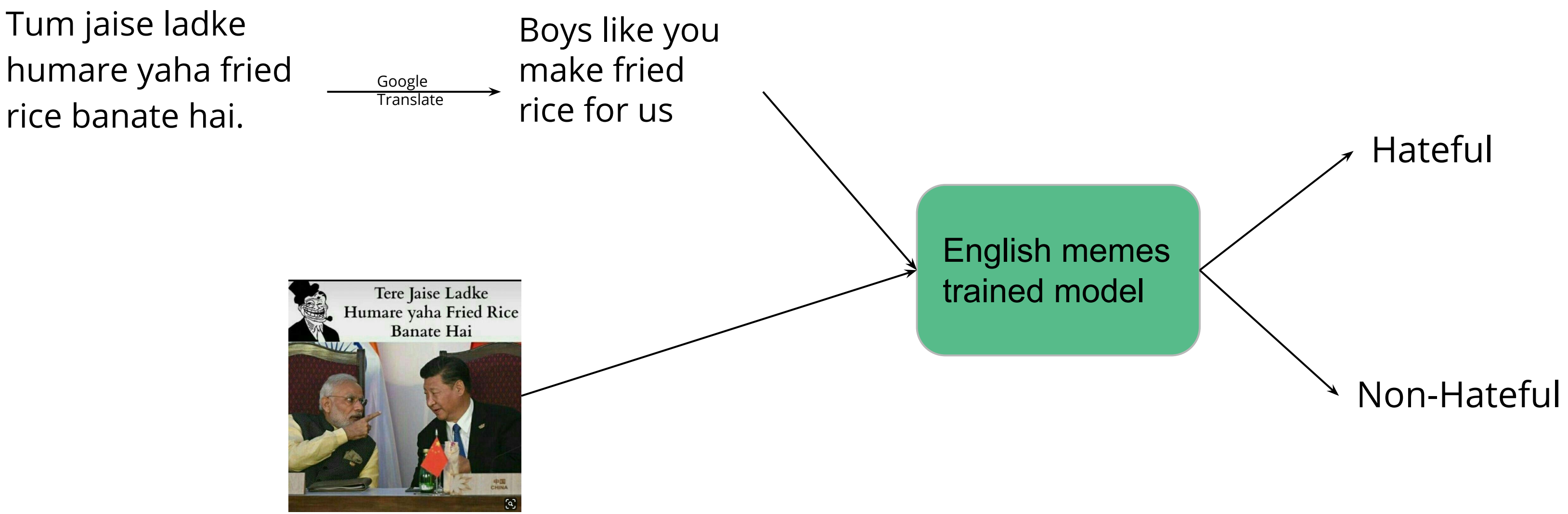
We applied the MOE ensemble on the models described in the VILIO approach. We got the above results, we can see that there is a slight improvement in the Dev Seen accuracy (0.4 %). Whereas, the accuracy on Test Seen decreases.
The reason for the decrease in accuracy is that, the performance of the individual models in the ensemble is almost same and hence the optimal alphas learnt by the MLP are same for all the models, hence the performance does not increase using MOE ensemble.

Beyond English memes



Can the same model classify this as hateful?

Pipeline



Tum jaise ladke humare yaha fried rice banate hai.

Google Translate

Boys like you make fried rice for us

English memes trained model

Hateful

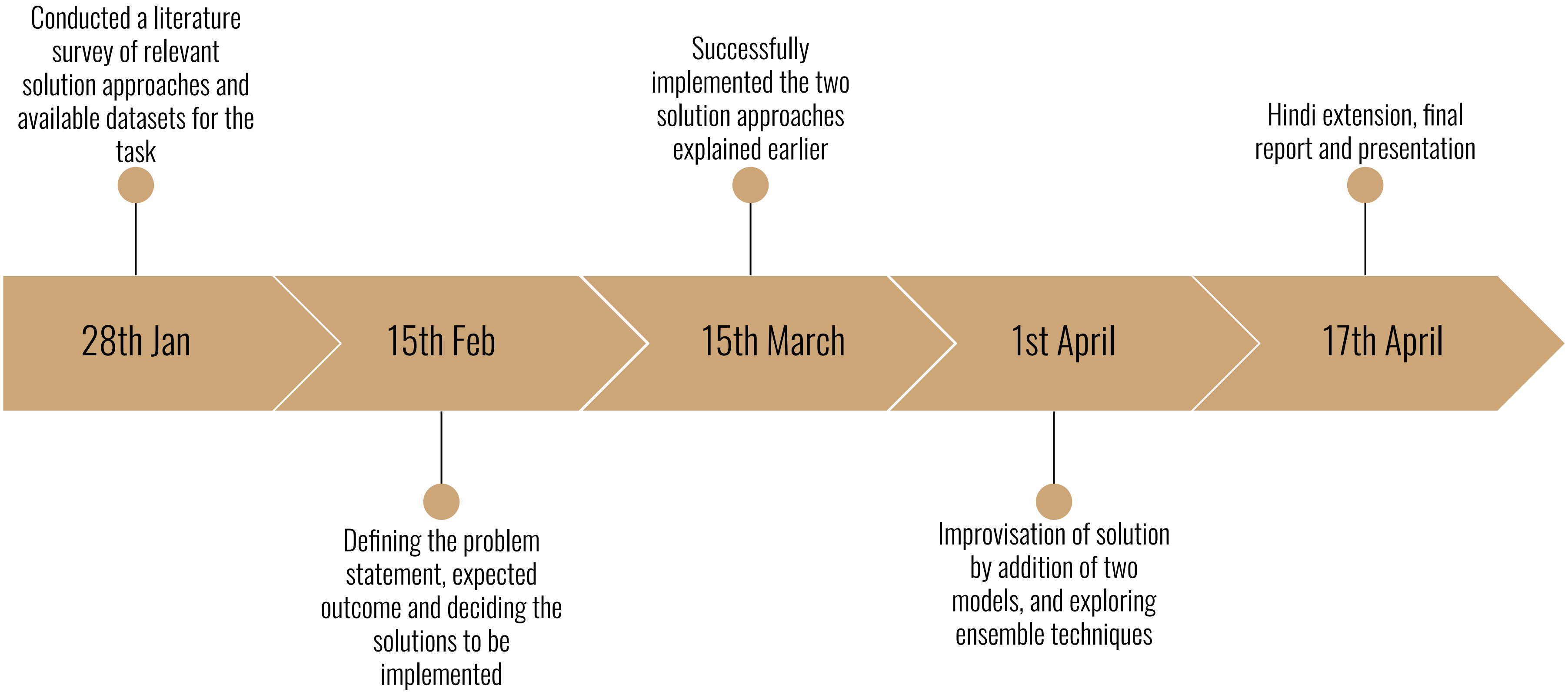
Non-Hateful

Results

Out of 39 total memes, only 22 were correctly classified(56.4%) by our model. Multiple reasons for poor performance:

- Pre-Training on large English datasets which do not contain Hindi references.
- Different kind of hateful content in the Facebook dataset vs. the memes we've collected.

Project Timeline



28th Jan

Conducted a literature survey of relevant solution approaches and available datasets for the task

15th Feb

Defining the problem statement, expected outcome and deciding the solutions to be implemented

15th March

Successfully implemented the two solution approaches explained earlier

1st April

Improvisation of solution by addition of two models, and exploring ensemble techniques

17th April

Hindi extension, final report and presentation

Special Thanks

Harika Abburi (TA Mentor)

QUESTION:
DO YOU HAVE ANY QUESTIONS?