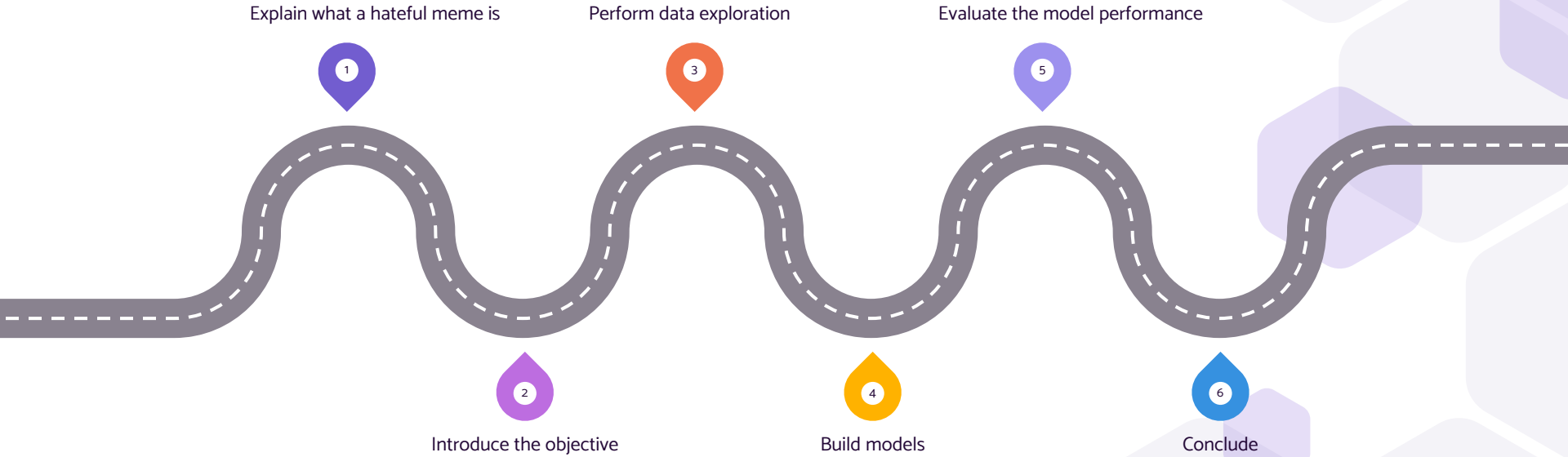
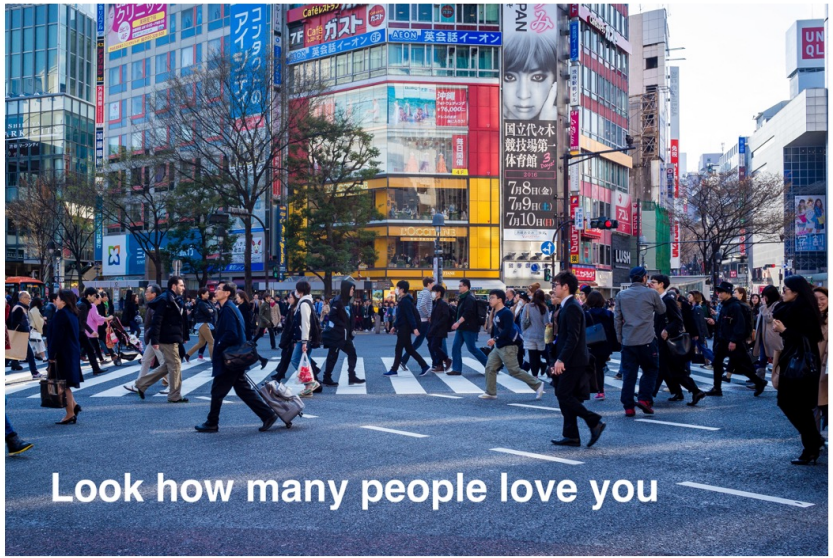
HATEFUL MEMES

Uyen Huynh

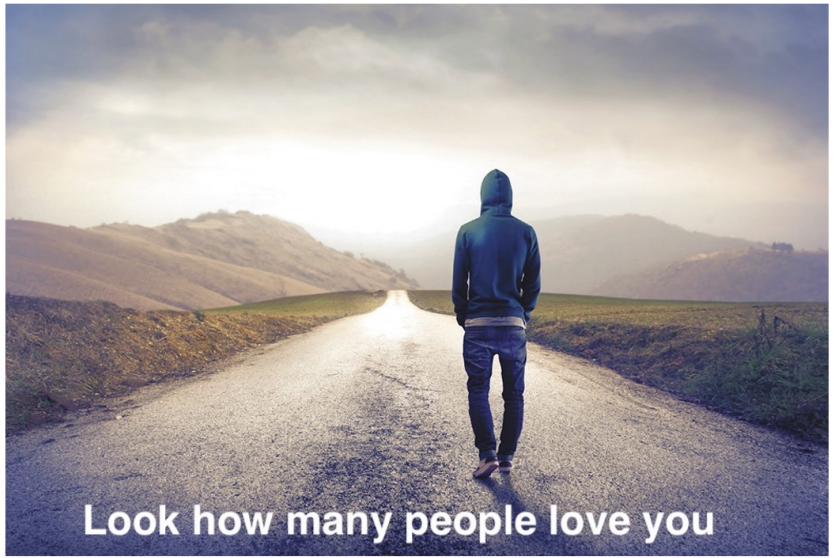
ROADMAP


Explain what a hateful meme is
1
Introduce the objective
2
Perform data exploration
3
Build models
4
Evaluate the model performance
5
Conclude
6

What is a hateful meme?

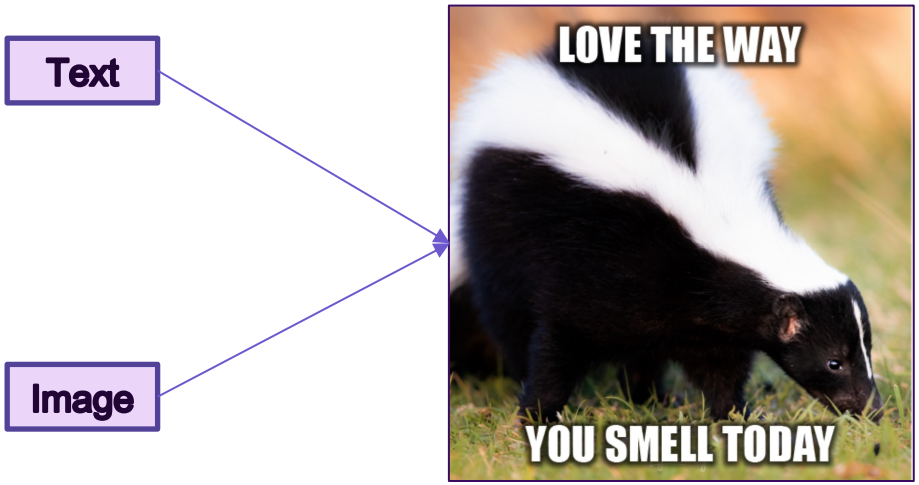


Class 0: Not hateful



Class 1: Hateful

What does a meme comprise of ?



To understand the true meaning of a meme, we can’t separately look at the image or text.

Objective

Design a multimodal convolutional neuron network to detect hateful speech in memes

Look how many people love you

Image features

+

Text features

Class 0: Not hateful

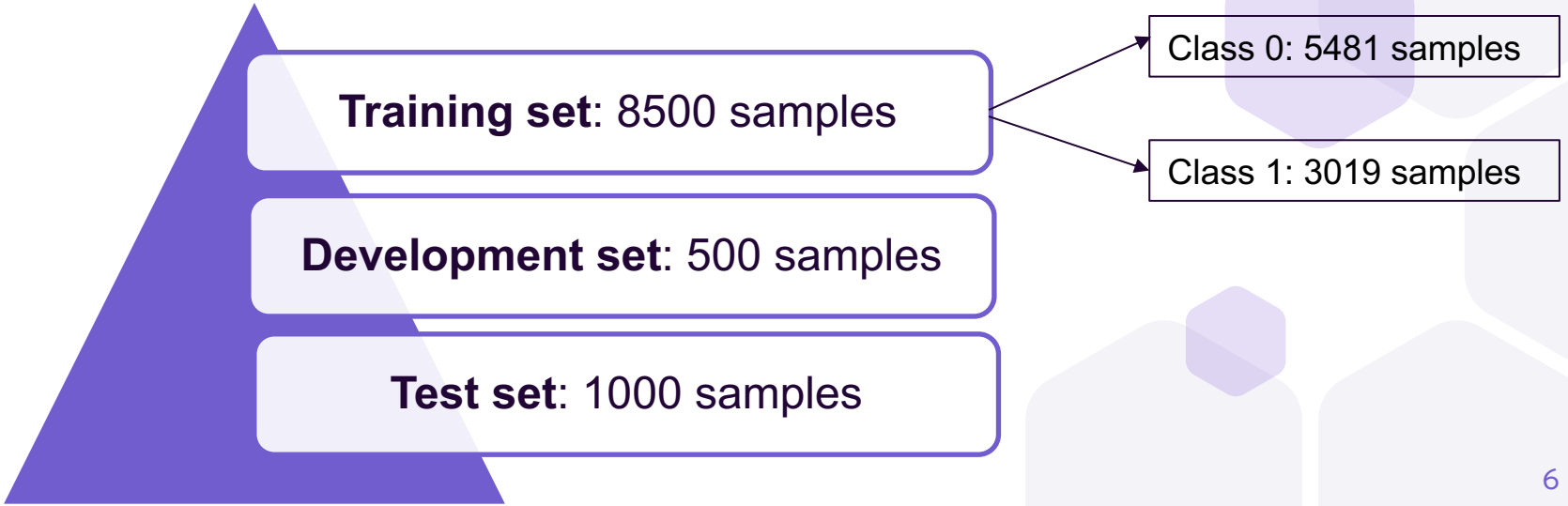
Class 1: Hateful

Input **Model** **Label**

Data

Source: The dataset is created by Facebook AI & can be downloaded on the *Hateful Memes Challenge* website.

The dataset contains pairs of text & image:



Data Preparation

Image

- Resize the images to **128 x 128** pixel.
- Normalize the pixel values using the mean and stddev of ImageNet

  mean = [0.485, 0.456, 0.406]

  stddev = [0.229, 0.224, 0.225]

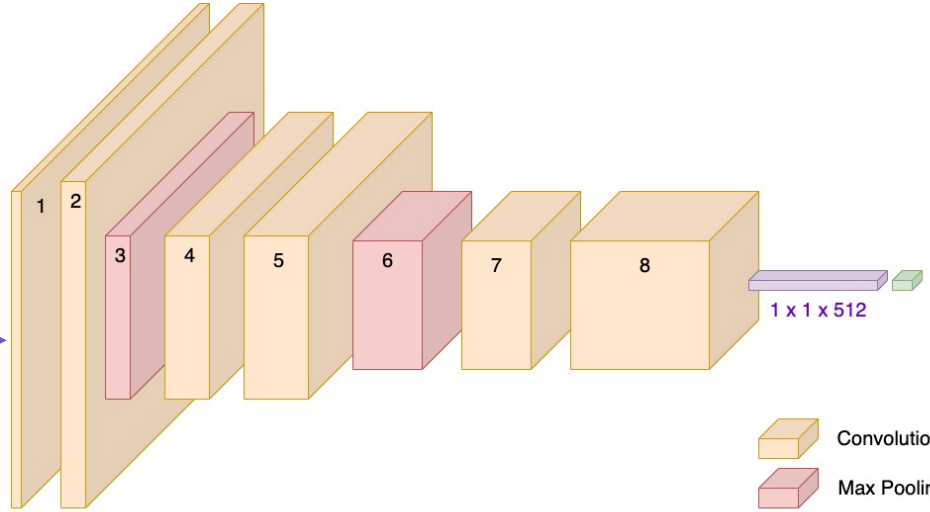
Text

- Using NLTK tokenize package to divide the sentence into a list of words.
- Remove stop words.

Model: Baseline Unimodal 1

Simple CNN: Using only images to predict the label

Input: 128x 128 x 3

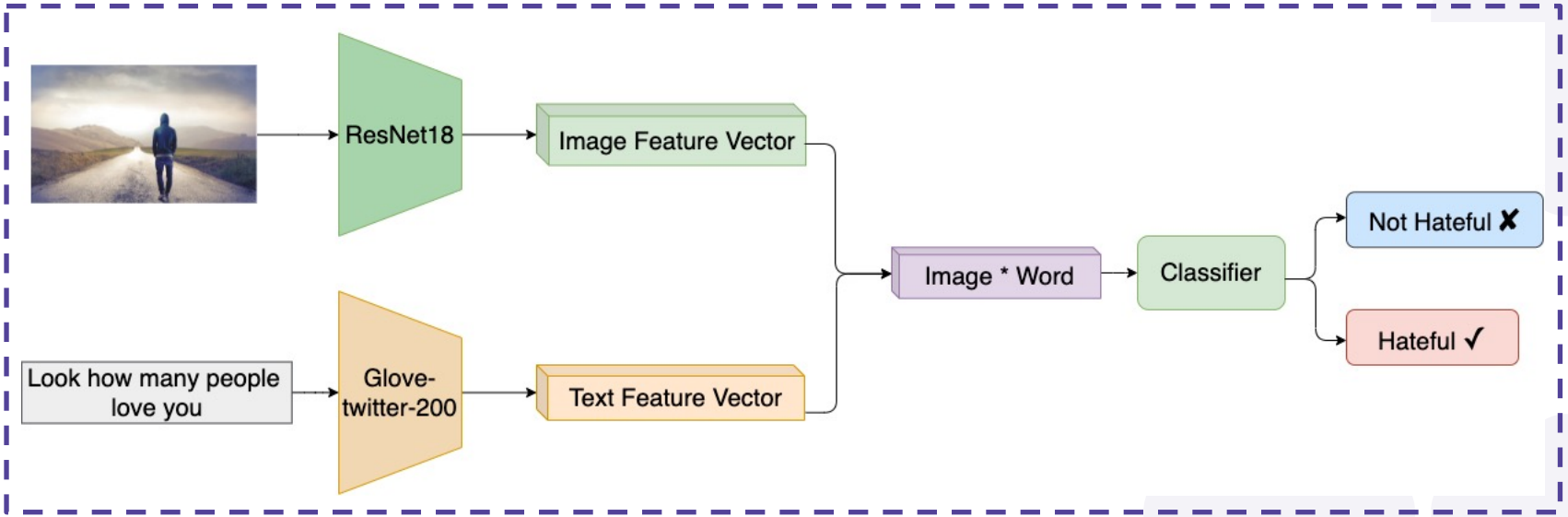


Model: Baseline Unimodal 2

Using pre-trained model ResNet18 to extract image features

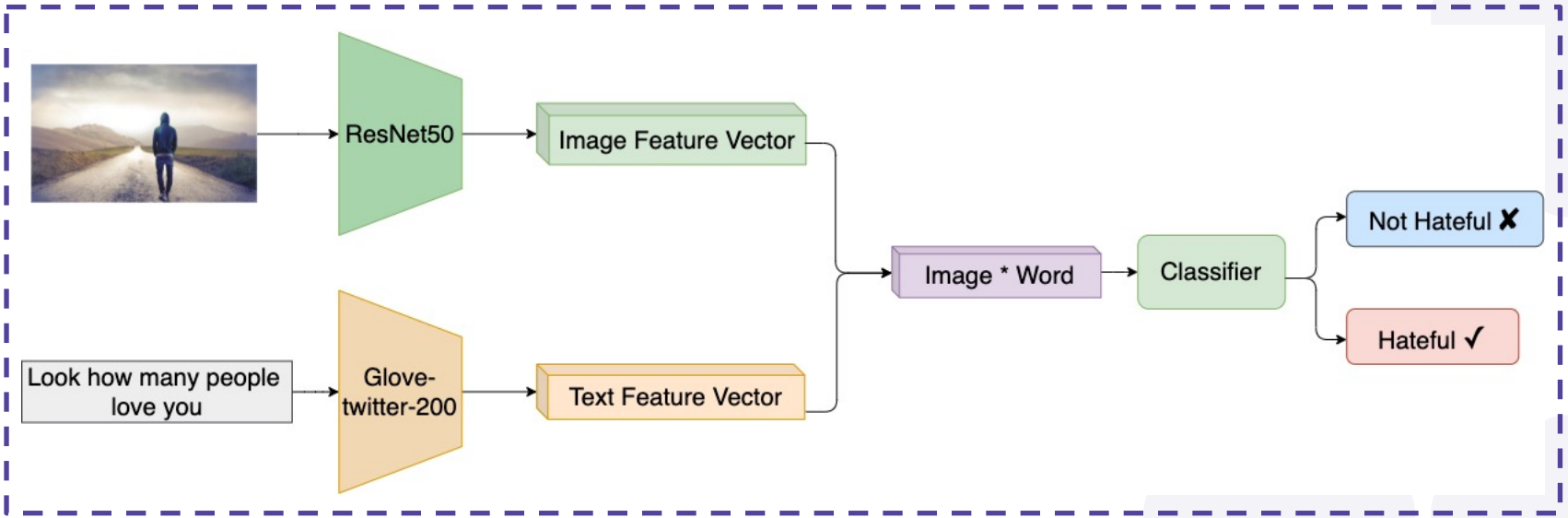
Model: Baseline Multimodal 1

Using pre-trained models: ResNet18 & glove-twitter-200



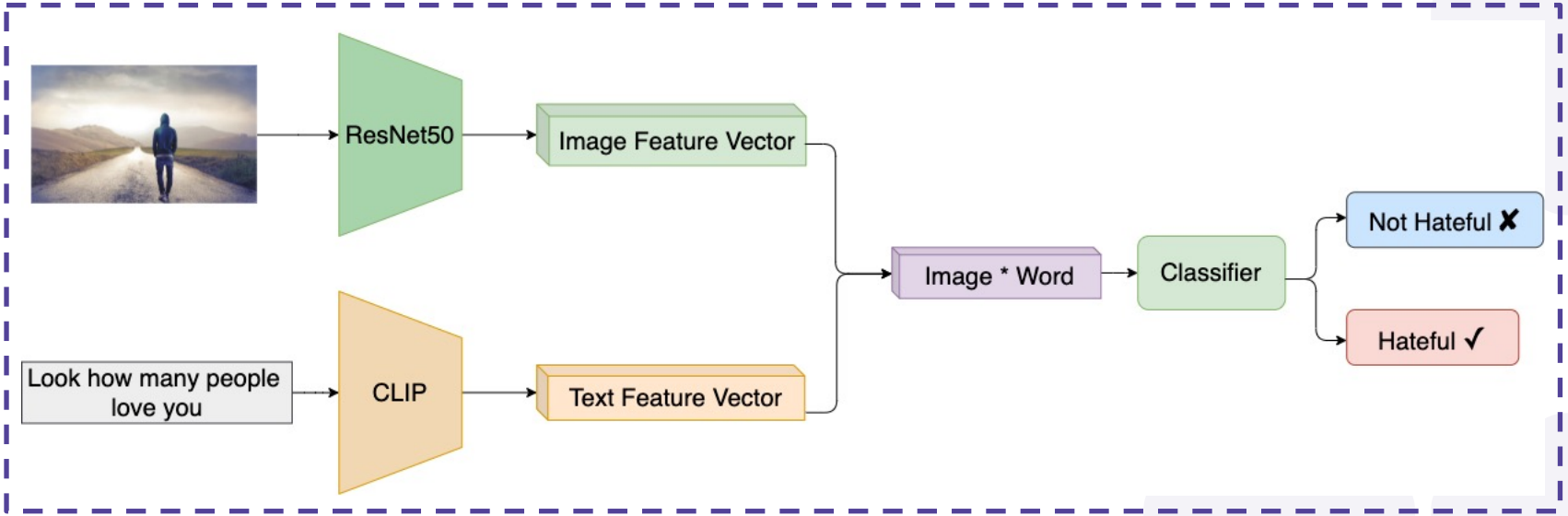
Model: Baseline Multimodal 2

Using pre-trained models: ResNet50 & glove-twitter-200



Model: Text Image Multimodal

Using pre-trained models: ResNet50 & CLIP



Evaluation Metrics

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

TP: True Positive
TN: True Negative
FP: False Positive
FN: False Negative

Model Evaluation

Model	Description	Development Accuracy
Baseline Unimodal 1	Designed CNN	0.506
Baseline Unimodal 2	ResNet18	0.524
Baseline Multimodal 1	ResNet18 & Glove-twitter-200	0.574
Baseline Multimodal 2	ResNet50 & Glove-twitter-200	0.576
Text Image Multimodal	ResNet50 & CLIP	0.616

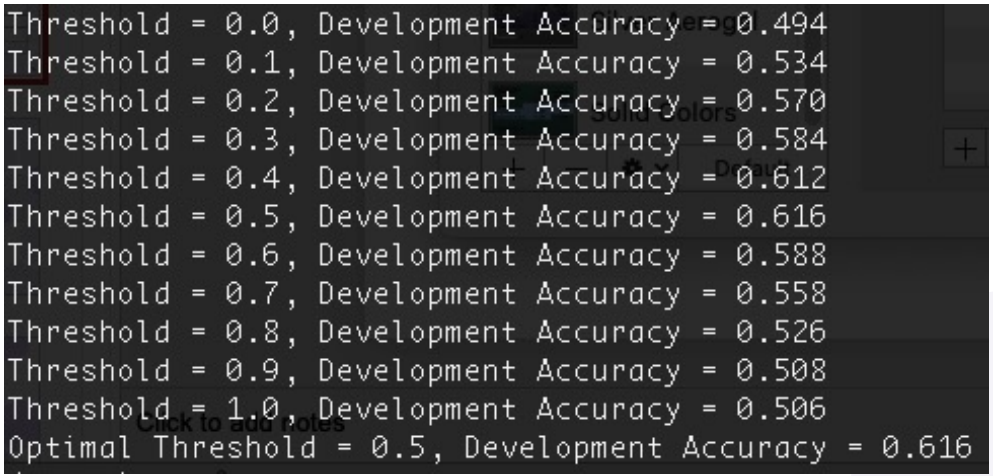
Text Image 3 has the highest development accuracy

Text Image 3: Optimal Threshold Selection

In the binary classification, the classifier returns a predicted label based on the threshold:

$$g(x) \gtrless \gamma \quad \begin{matrix} \textit{Hateful} \\ \textit{Not Hateful} \end{matrix}$$

Let's choose the optimal threshold for Text Image 3



Text Image 3: Model Performance

Test Accuracy = 0.637

Confusion Matrix

		Predicted	
		Class 0	Class 1
Actual	Class 0	TN = 429	FP = 81
	Class 1	FN = 282	TP = 208

Result Comparison

Type	Model	Validation Accuracy	Test Accuracy
Unimodal	Human		84.70
	Image-Grid	50.67	52.73
	Image-Region	52.53	52.36
	Text BERT	58.27	62.80
Multimodal (Unimodal Pretraining)	Late Fusion	59.39	63.20
	Concat BERT	59.32	61.53
	MMBT-Grid	59.59	62.83
	Text Image	61.60	63.70
	MMBT-Region	64.75	67.66
	ViLBERT	63.16	65.27
	Visual BERT	65.01	66.67

Source: The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes, FacebookAI

Conclusion

The **Text Image** Model is good at detecting hate speech related to:

- Religion
- Race
- Gender
- Terrorism