Multimodal Meme Classification

Eunbin S. Kim, Yiyang Liu
Erdos Institute Deep Learning Bootcamp

The TASK: To identify the sarcasm in Memes

WHAT THE DUCK
THIS IS A MEME
MOM, I JUST NEED 2 FOCUS ON MY ART
Regular
Sarcastic

Binary classification

Sarcastic vs. not sarcastic

Input: A Meme

[Captions]
[Images]

Output: Class Labels.

[A regular meme] = 0
[A sarcastic meme] = 1

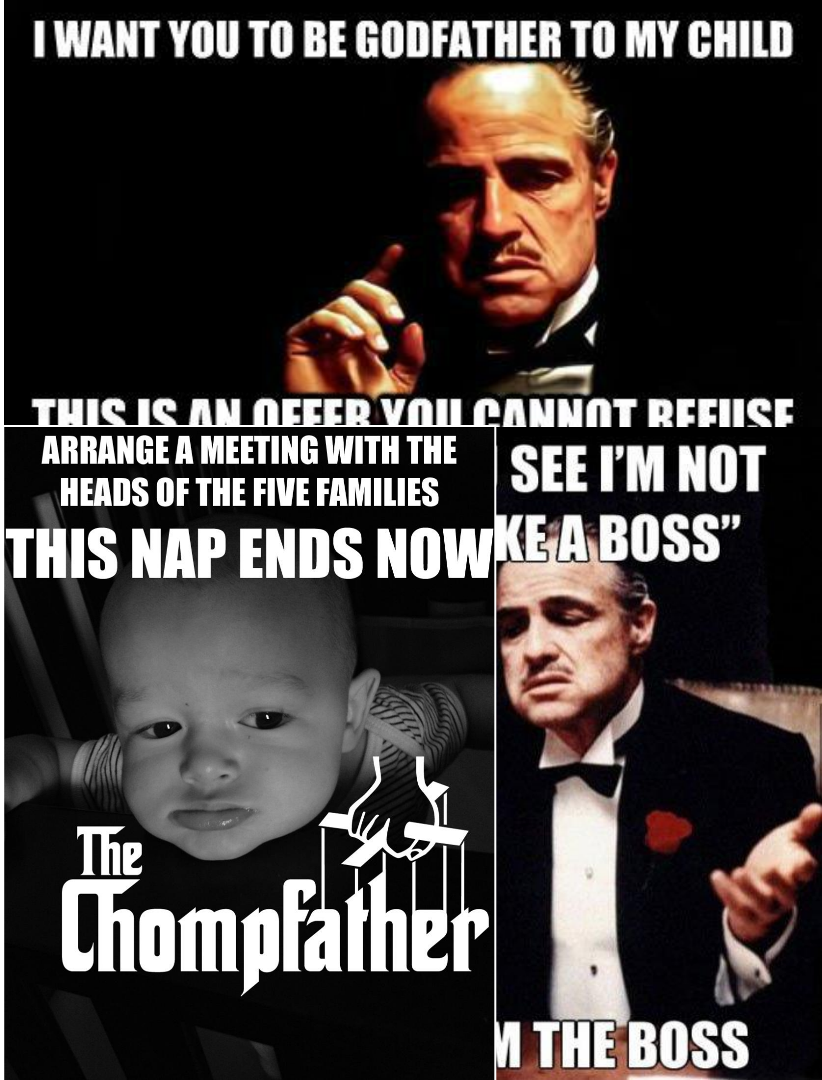
Multimodal model

Unimodal models: one type of data

Multimodal models: combine multiple data types

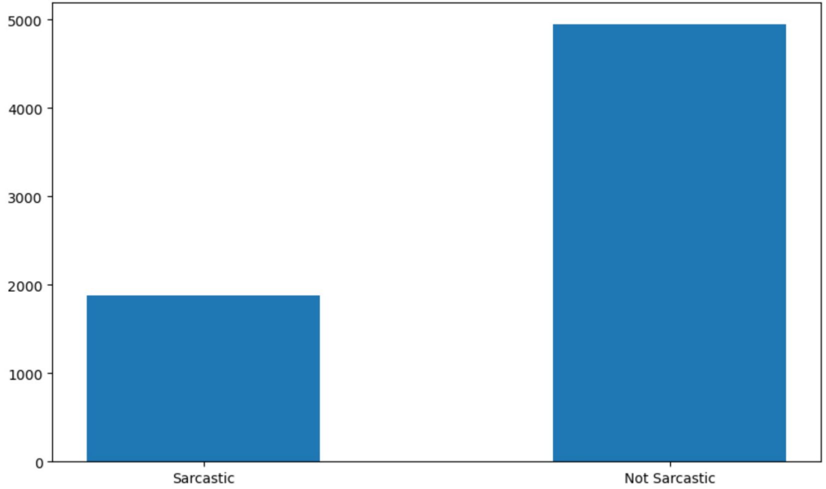
The sarcasm in memes sometimes lie in the contrast between the images and captions.

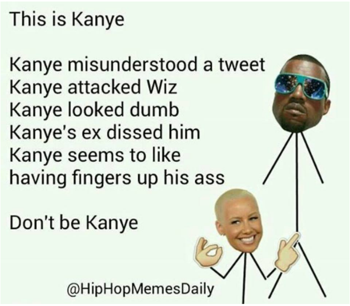
- Godfather + Marlon Brando = NOT TWISTED
- Godfather + Toddler = TWISTED


I WANT YOU TO BE GODFATHER TO MY CHILD
THIS IS AN OFFER YOU CANNOT REFUSE
ARRANGE A MEETING WITH THE HEADS OF THE FIVE FAMILIES
THIS NAP ENDS NOW
The Chompfather
SEE I'M NOT
KE A BOSS"
M THE BOSS

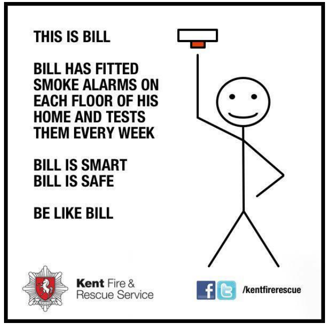
The DataSet

- Features: Images, Caption, Labels
- The total # of memes: 7000 (6830 after cleaning)
- Sarcastic: 1884 Not Sarcastic: 4946
  - ○ Ratio: (sarcastic) 1: 2.78 (not sarcastic)
  - ○ Slightly imbalance → metrics: auc-roc





Sarcastic



not sarcastic

PRE-PROCESSING + ENCODING

Captions pre-processing:

- Strip all special characters, remove watermarks.
- *Lemmatization*
- Remove stop words

TEXT ENCODER: DISTILBERT

Image pre-processing:

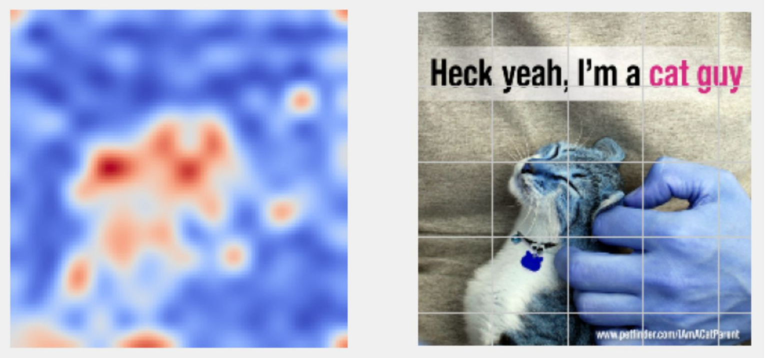
- Clean corrupted images, convert to RGB space.
- Normalize the image

IMAGE ENCODER: VIT DINOv2

I PREFER THE REAL ZELDA! I SAID THE REAL ZELDA! PERFECTION imgflip.com

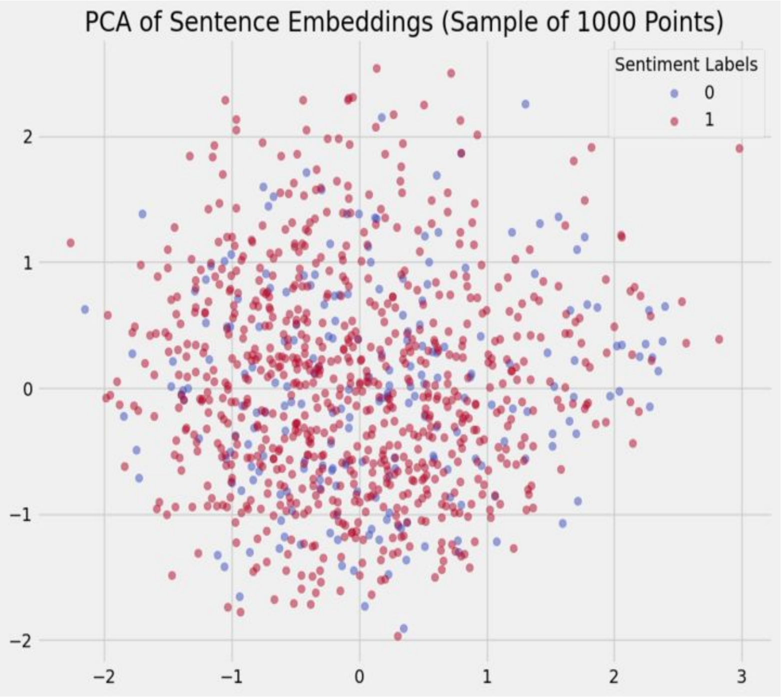
I prefer the real zelda! I *say* the real zelda! *perfect* imgflip.com

ATTENTION! It's a CAT



Visualized Attention vs Image.

Visualized embedding



Sentence Embedding (DISTILBERT)

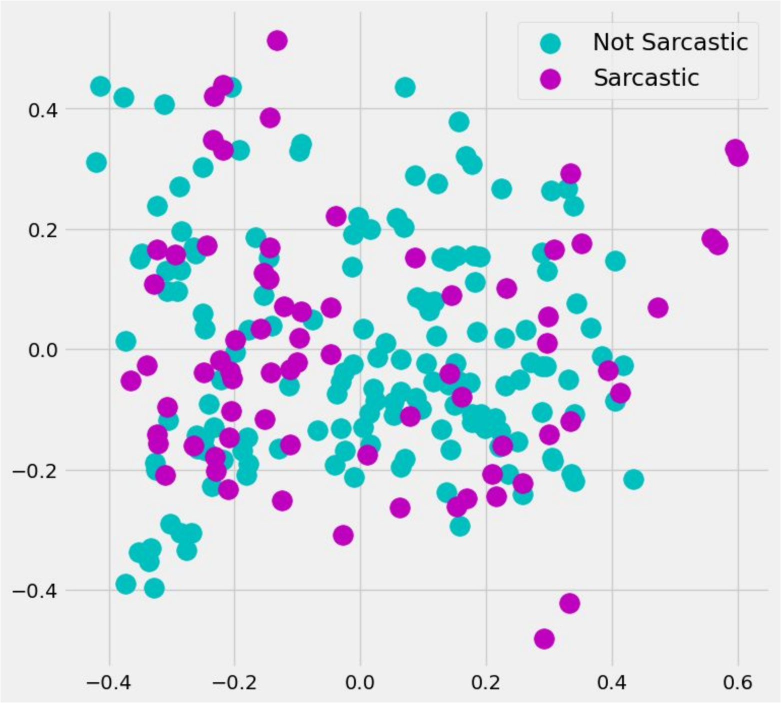
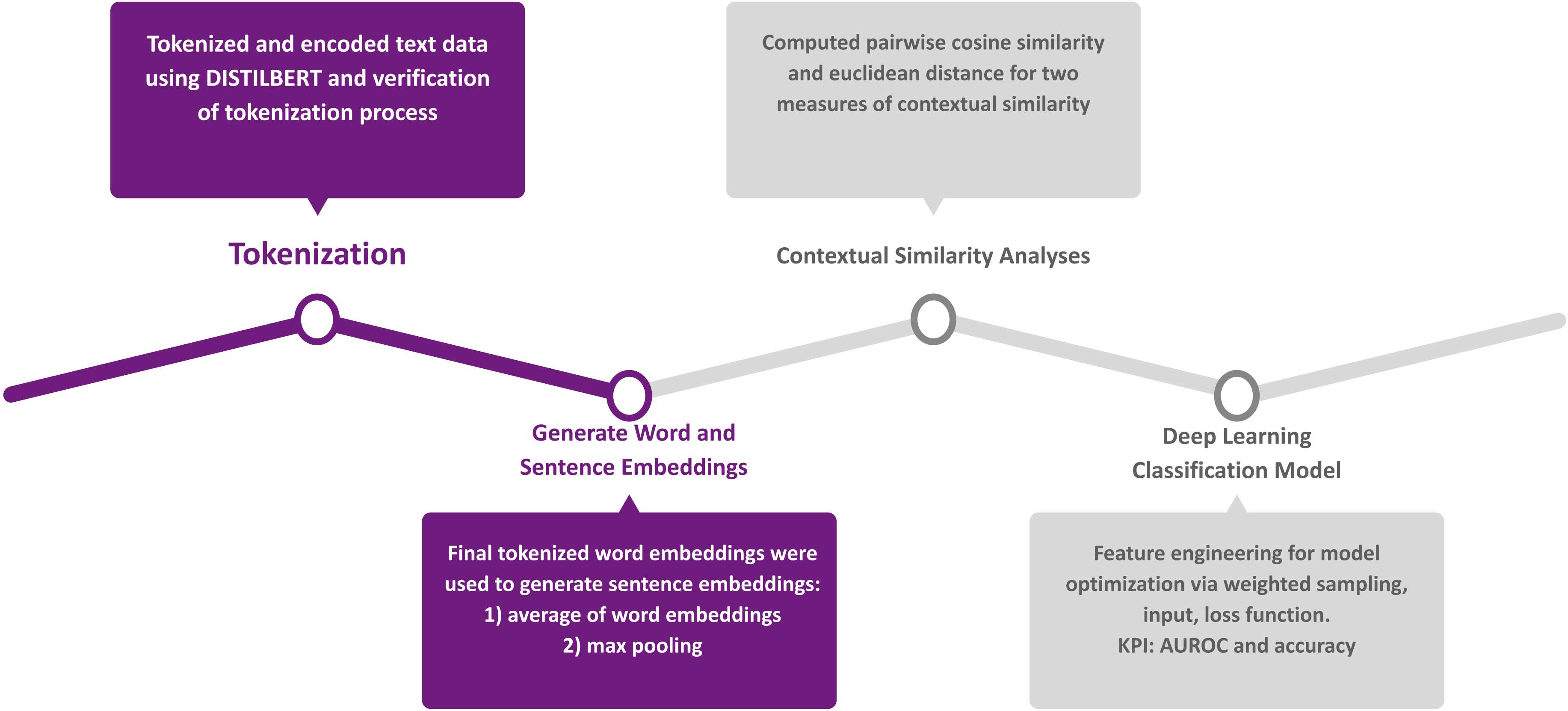


Image Embedding (DINOv2)

Exploratory NLP Data Analysis (DistilBERT)



Tokenization

Tokenized and encoded text data using DISTILBERT and verification of tokenization process

Generate Word and Sentence Embeddings

Final tokenized word embeddings were used to generate sentence embeddings:
1) average of word embeddings
2) max pooling

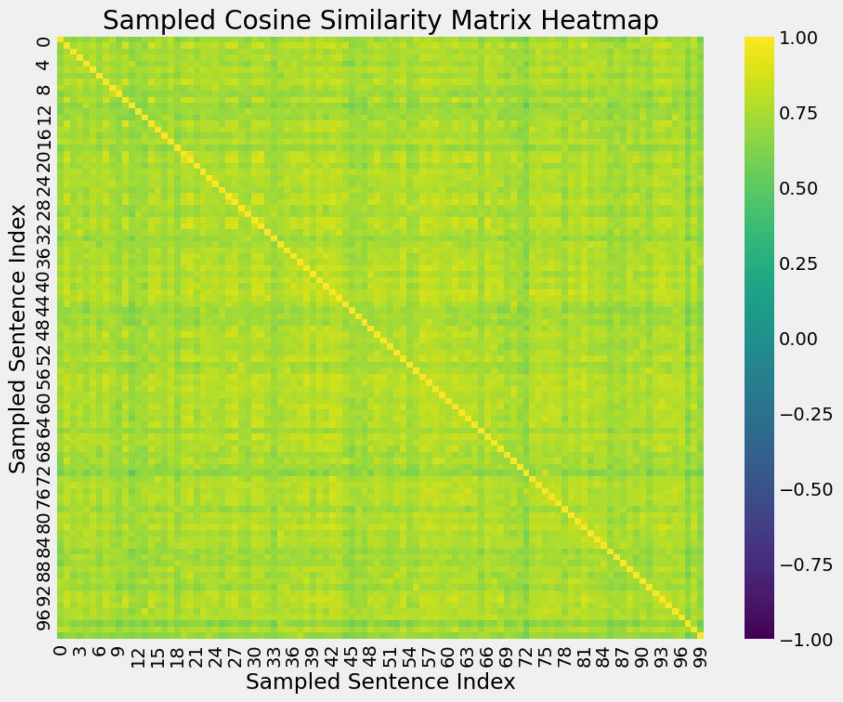
Contextual Similarity Analyses

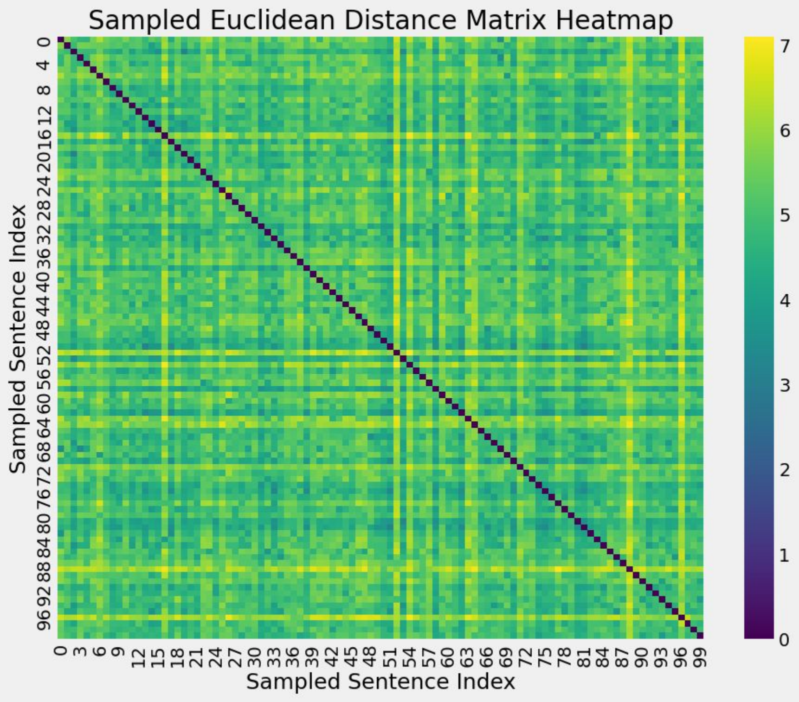
Computed pairwise cosine similarity and euclidean distance for two measures of contextual similarity

Deep Learning Classification Model

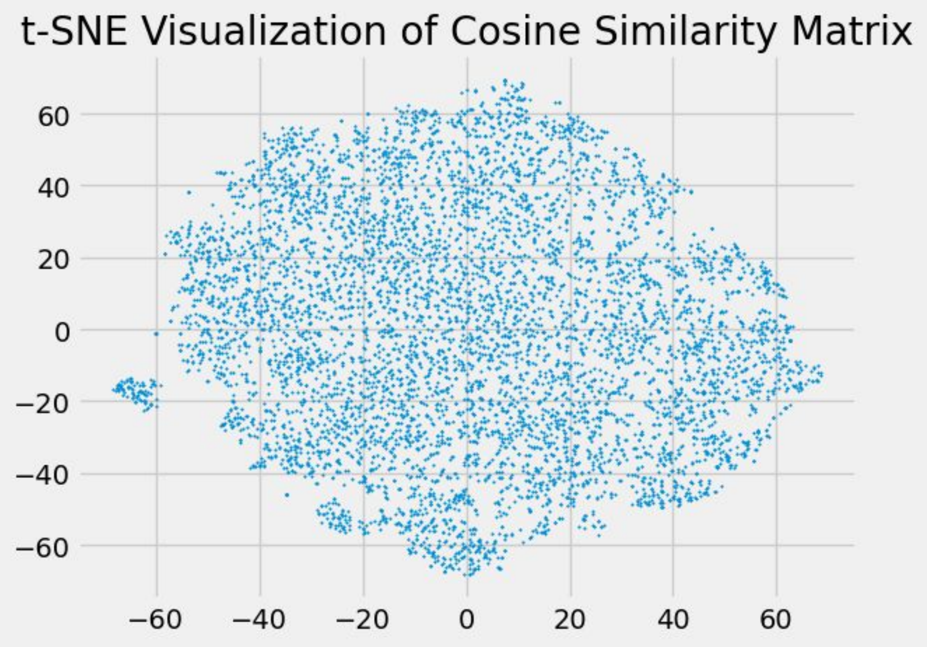
Feature engineering for model optimization via weighted sampling, input, loss function.
KPI: AUROC and accuracy

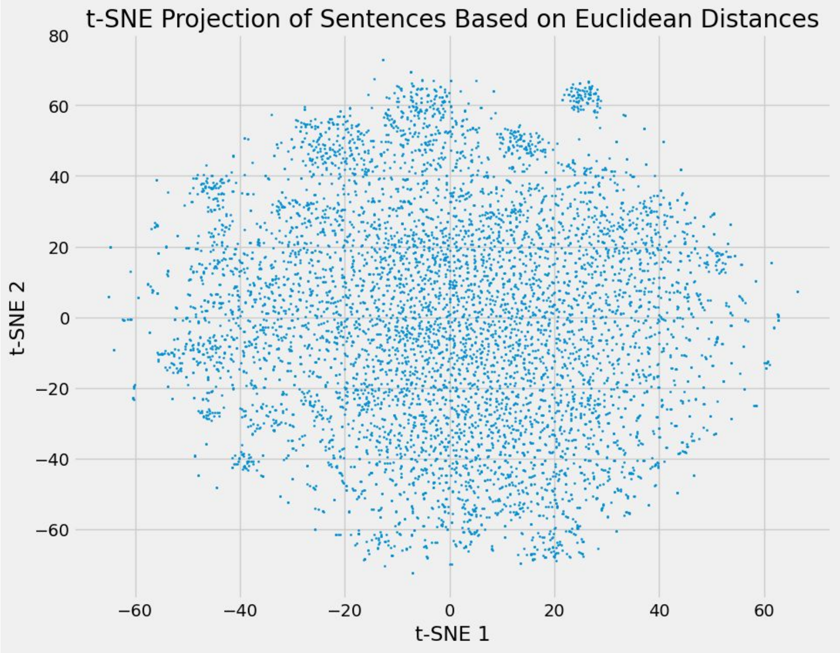
Contextual similarity Comparison





Contextual similarity Comparison





THE MODEL

reference:Hate-CLIPper

Components in the architecture:

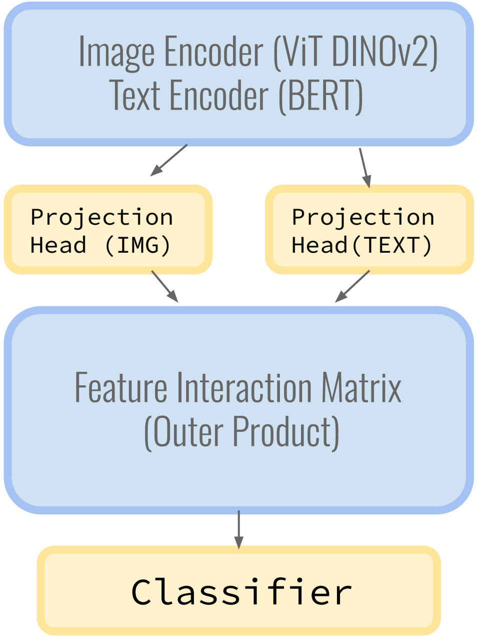


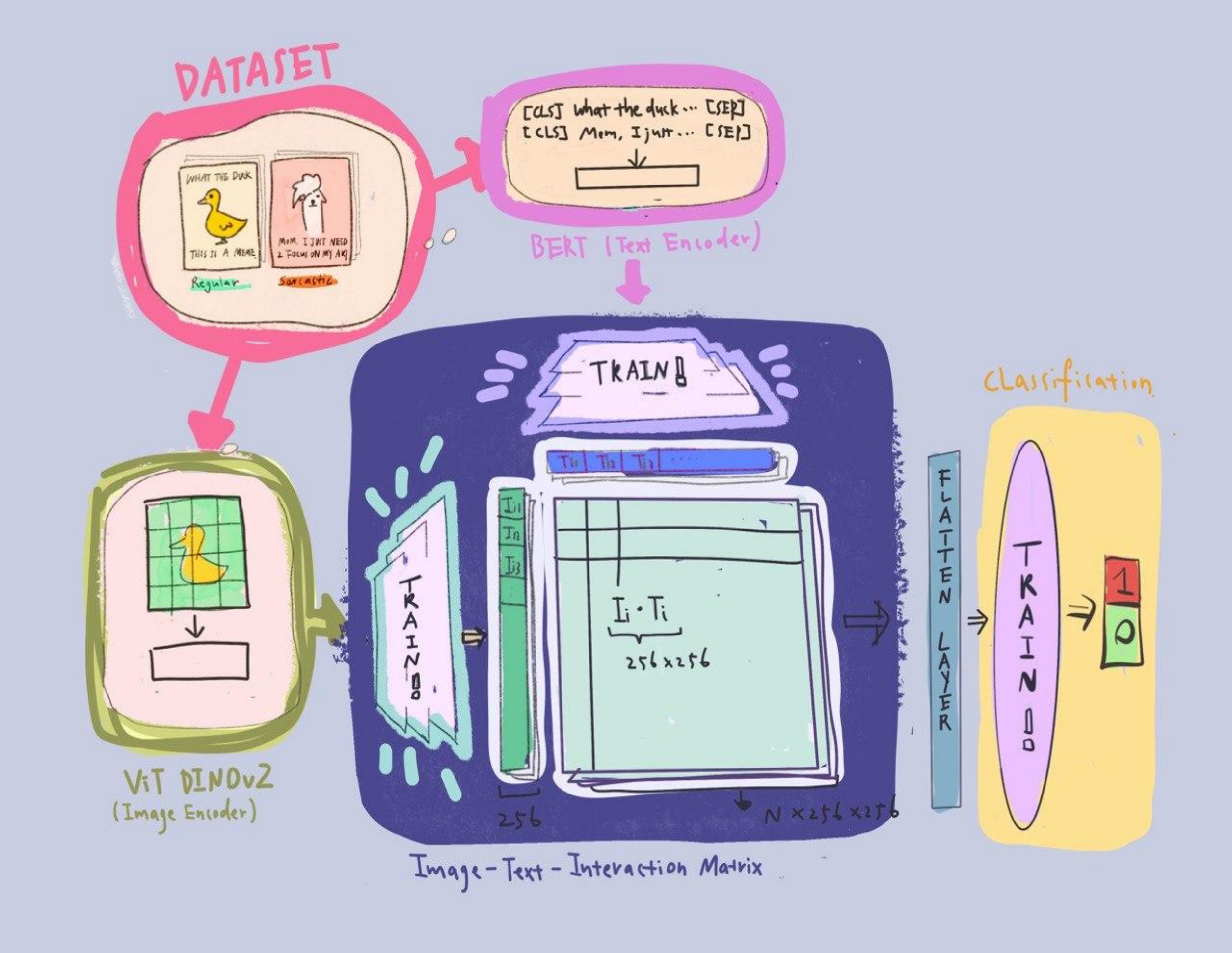
Image Encoder (ViT DINOv2)
Text Encoder (BERT)

Projection Head (IMG)

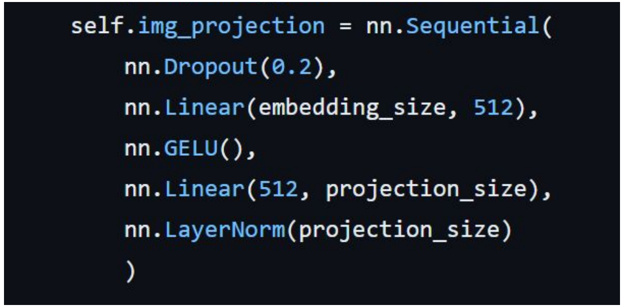
Projection Head(TEXT)

Feature Interaction Matrix
(Outer Product)

Classifier

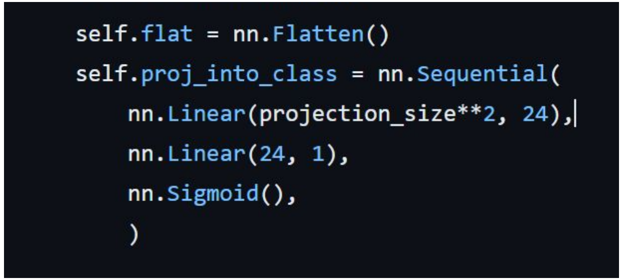


KEY IMPLEMENTATION (FOR REFERENCE)



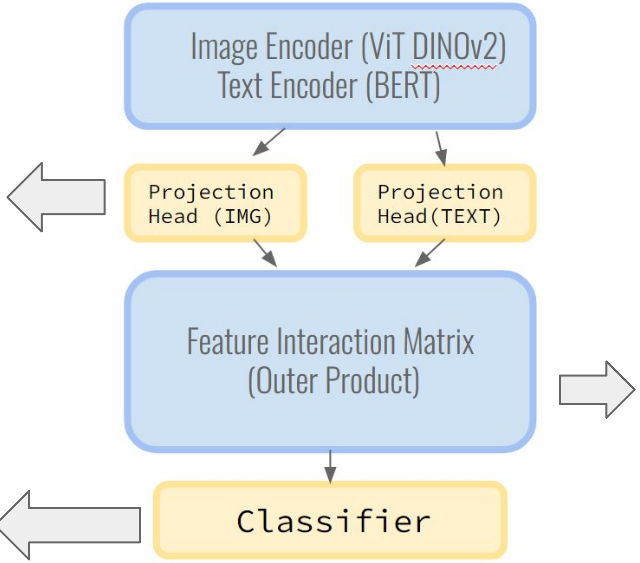
```
self.img_projection = nn.Sequential(
    nn.Dropout(0.2),
    nn.Linear(embedding_size, 512),
    nn.GELU(),
    nn.Linear(512, projection_size),
    nn.LayerNorm(projection_size)
    )
```

Projection heads



```
self.flat = nn.Flatten()
self.proj_into_class = nn.Sequential(
    nn.Linear(projection_size**2, 24),
    nn.Linear(24, 1),
    nn.Sigmoid(),
    )
```

Classifier



Img_feature: I_i
Batch_size x 1 x 256

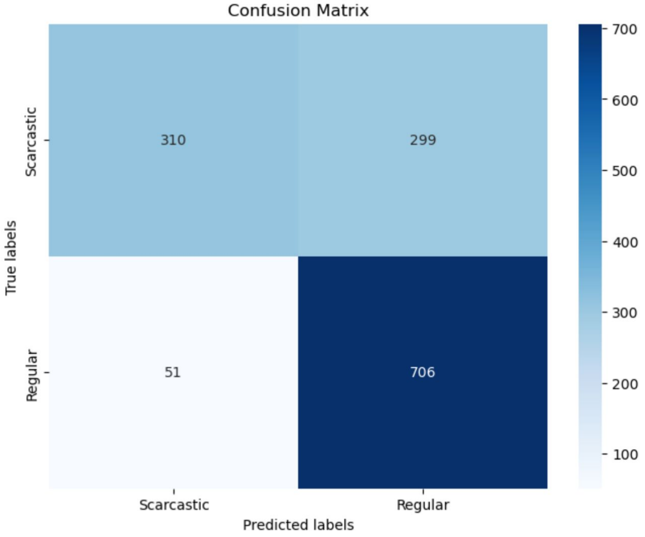
Text_feature: T_i
Batch_size x 1 x 256

Feature interaction:
FIT_i = I_i.T @ T_i
Batch_size x 256 x 256

Classification Results

Multimodal Model:

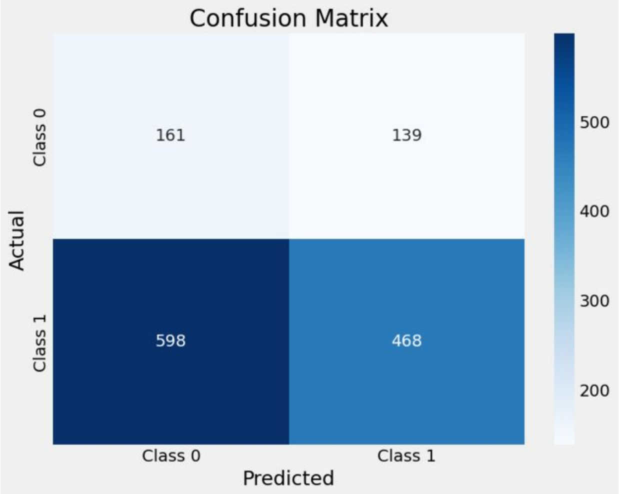
- Accuracy: 0.7436
- **AUC-ROC: 0.7969**



Comparison:

Uni-modal Model (DistilBERT)

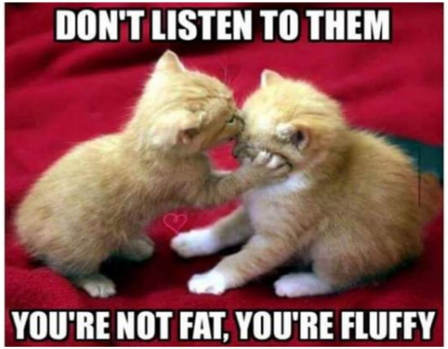
- Accuracy: 0.4605
- **AUC-ROC: 0.4868**



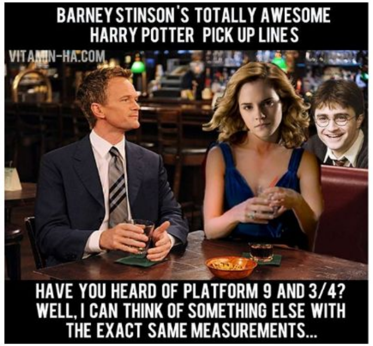
KPI and Conclusions

- **KPIs: Accuracy and AUROC**
- **The multimodal model outperformed the unimodal model**
- **Sarcasm in memes involves a blend of visual humor and textual irony**
- **Benefits of the multimodal model:**
  - Enhanced understanding: able to better capture the nuance of sarcasm in memes with both text and visual cues
  - Improved Accuracy: Integrating image and text data typically leads to better performance than using only text, as sarcasm often relies on both visual and linguistic features
- **Text only models might miss contextual cues crucial for accurate classification provided by the image**

False Positive



False Negative


BARNEY STINSON'S TOTALLY AWESOME HARRY POTTER PICK UP LINES
VITAMIN-HA.COM
HAVE YOU HEARD OF PLATFORM 9 AND 3/4? WELL, I CAN THINK OF SOMETHING ELSE WITH THE EXACT SAME MEASUREMENTS...

Thank you all for listening!

Special thanks to:

Erdos Institute
Lindsay Warrenberg
Marcos Ortiz

THE ERDŐS INSTITUTE
Helping PhDs get and create jobs they love at every stage of their career.