Social Computing Group,
School of Computation,
Information and Technology,
Technical University of Munich

International Workshop on Semantic Evaluation

Detection of Persuasion Techniques in Memes

Gözde Ünver, Batikan Özkul, Fami Mahmud and Daryna Dementieva

Motivation and Goals

- Usage of memes for spreading disinformation and propaganda on social media is amplifying
- Automatic detection of persuasive content is becoming increasingly important

SemEval 2024 challenge Task 4 problem definitions:

- Subtask 1: Hierarchical multilabel classification using only textual content of memes (20 persuasion techniques)
- Subtask 2A: Hierarchical multilabel classification on both images and texts together (22 persuasion techniques)
- Subtask 2B: Binary classification on both images and texts together
- 32 research teams participated in this challenge

[1]

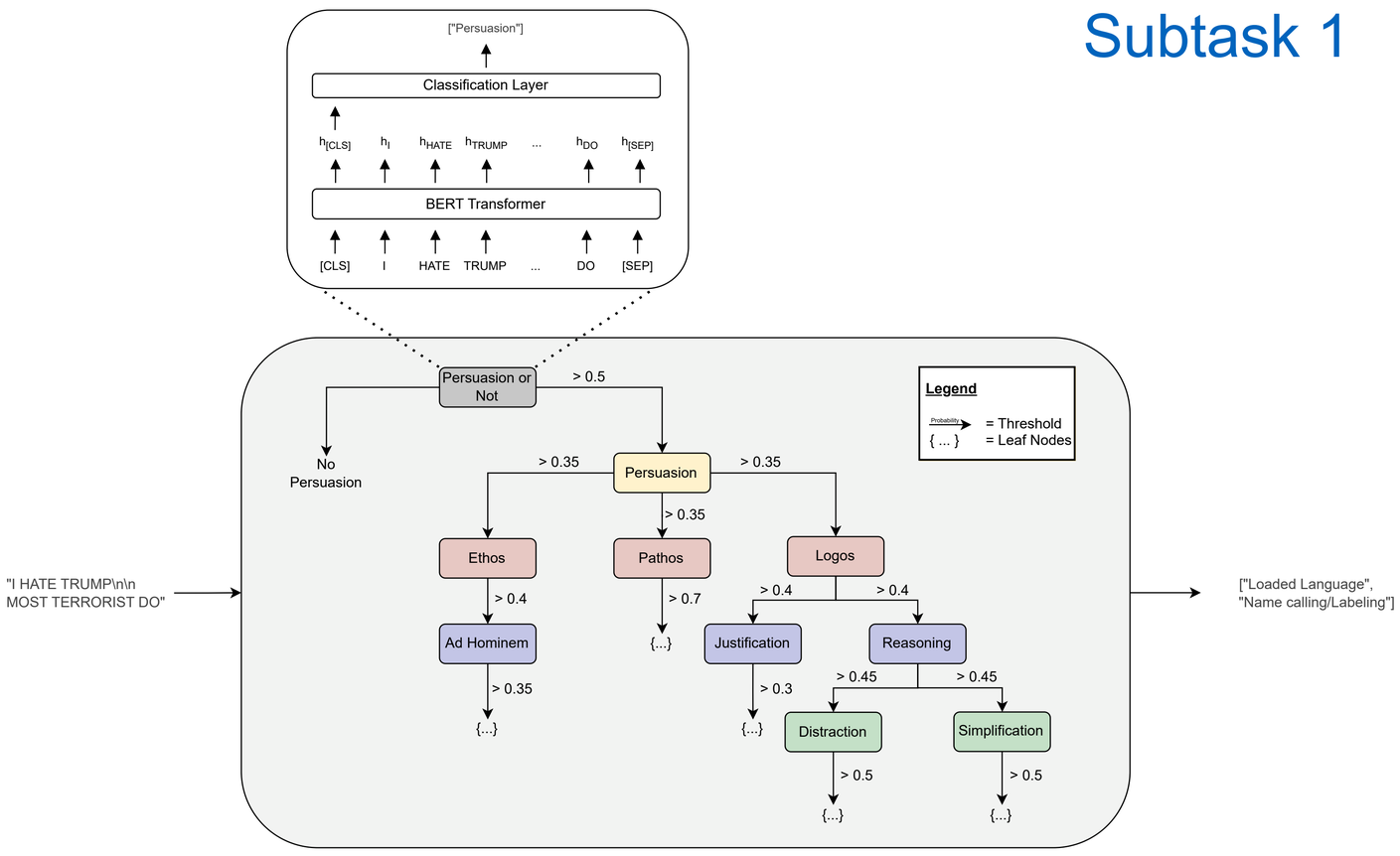

```
{
        "id": "125",
        "text": "I HATE TRUMP\n\nMOST TERRORIST DO",
        "labels": [
                "Loaded Language",
                "Name calling/Labeling"
],
"link": "https://..."
},
```


[2]

- All memes are in English
- Each meme contains a text in the image
- Text is also present in json files

	Train Set	Validation Set	Development Set	Test Set
Subtask 1	7000	500	1000	1500
Subtask 2A	7000	500	1000	1500
Subtask 2B	1200	150	300	600

Subtask 1



Hierarchical Multilabel Classification

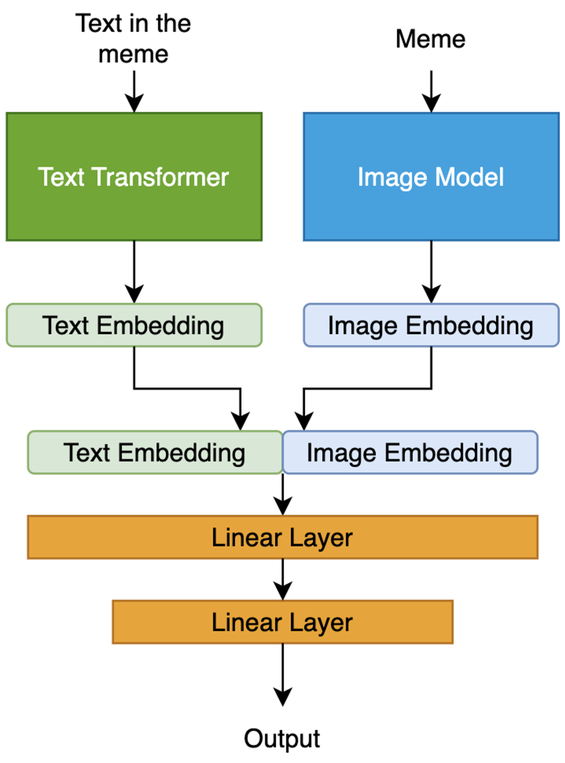
- Fine-tuned model in each node of hierarchy predicts child nodes
- Text-Transformer: DeBERTa-V3-large
- Iterative hierarchical processing: Samples passed from parent to child node if child node prediction probability exceeds threshold
- Final multilabel prediction when all child node probabilities are below threshold or leaf nodes are reached

Evaluation

- Hierarchical reward system based on F_1-score
- Full reward for exact leaf node predictions
- Partial reward for predictions matching an ancestor of the correct leaf node

Weights & Biases · **SemEval** · Hugging Face · PyTorch

Subtask 2



Subtask 2A

- End-to-end training by concatenating embeddings of text and image as in Subtask 2B
- Multilabel classification of persuasion techniques
- Text transformer: BERTweet-large with optimized threshold of Subtask 1
- Due to time constraints, only tested by the best performing models from the previous tasks.
- Evaluation using hierarchical-F_1

Subtask 2B

- End-to-end training by concatenating embeddings of text and image
- Text transformer: BERTweet-large
- Experimented with many image models, including CNN-based models and vision transformers, best performing; google/vit-base-patch32-224-in21k
- Experimented with different embedding methods: CLS, pooler_output, the average of all tokens
- Used the average of all tokens method as it was the best performing one on the development set
- Evaluation using macro-F_1

Further Experiments

Subtask 1

- Transfer learning: Child node models initialized with parent node's fine-tuned model weights
- Ensembling multiple models using stacking method with random forest as classifier
- Few-shot classification using GPT-4 and Llama

Subtask 2B

- We tried out cross attention, ensembling image and text models using random forest classifier and also linear layers
- Testing the model performance on the updated dataset after removing the texts from the images with a pre-trained Keras-OCR model

Results

	F1 Dev	F1 Test	Ranking Dev	Ranking Test
Subtask 1	0.63918	0.67384	12	4
Subtask 2A	0.67846	0.67717	5	6
Subtask 2B	0.85366	0.78413	1	9

F1 scores in this table refer to F1-hierarchical for Subtask 1 & 2A and F1-macro for Subtask 2B

[1,2]: https://propaganda.math.unipd.it/semeval2024task4/index.html#organisation