Improving Hateful Memes Identification with Ensemble Learning

HENG Theo, SABATIER Aubin, VAN DEN HEUVEL Victor

Group 31

Problem definition

The aim of this project is to study a new solution for detecting hateful content in a meme using an ensemble learning structure. In order to improve the performance of current models, several models, each trained on a specific topic, are stacked to classify hateful memes from non-hateful memes. It is therefore possible to compensate for the weaknesses of each model.

Key Related Works

Although Ensemble Learning structures are not new, the models are generally trained to correspond to specific file types and not to particular topics. The models must therefore learn to recognise several forms of hate content, directed at several different communities [1]. The aim is therefore to propose a different way of training models.

Dataset(s)

The dataset used is taken from the dataset proposed for the Hateful Memes Challenge competition organised by the Meta group. This is an open source dataset designed to measure progress in multimodal classification of vision and language. [3]
More precisely, the memes were first grouped by class using the keywords appearing in the memes' description, and then the 3 most represented classes were selected.

Method

1. Dataset formatting using keywords

Clases	Size	"Africans"	"Womens"	"Muslims"	Harm rate
Dataset A	1186	360	0	0	25.63%
Dataset W	1372	0	241	0	9.23%
Dataset M	1312	0	0	362	17.00%
Dataset Base	3454	518	352	448	×

Table 1: Dataset structure for fine tuning the models

2. **Choice of models:**

The stacking structure is applied to 3 state-of-the-art models with different architectures to ensure that the results obtained did not depend solely on the type of model: CLIP, BLIP, ALIGN

3. **Training the models:**

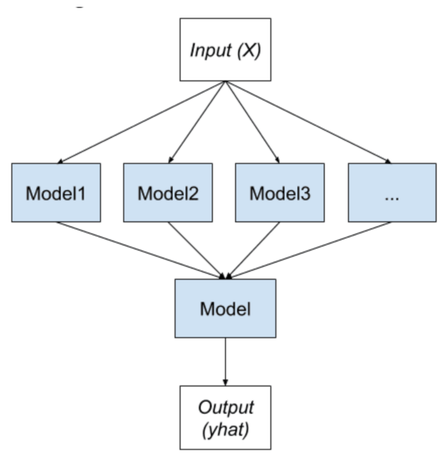
Fine-tuning : use of pretrained models to start from already extracted feature

Hyperparameters	CLIP	BLIP	ALIGN
Dataset A	5e-5	5e-6	5e-5
Dataset W	32	32	32

EE-559: Deep Learning, 2024

Limitations

This method involves parallelizing multiple models to compensate for each other's weaknesses. In the first part of the structure, each model provides an estimated label for the input meme, The estimates are then reconciled from each model by trusting the model with the highest estimate. [2]



Validation

1. Testing the results:

Model	With EL	Without EL	Results	Model A	Model W	Model M
CLIP	63.86%	61.61%	✓	57.8%	47.19%	56.22%
BLIP	40.18%	49.50%	×	41.73%	46.05%	49.62%
ALIGN	-	-	-	54.65%	-	51.32%

Table 2: Models performance based on F1-score

We can see that modelA, modelW and modelM are not performing great alone because they are not trained to detect hate outside of their class. However, by combining the predictions of the three models, it is possible to improve the performance.

2. Analysis:
   - The ensemble learning method can bring better results but it is not the case for all of the models
   - CLIP shows improvements, in contrast to BLIP
   - We noticed Training and testing depend notably on the random initialization of the training

Limitations

The main limitation of this method is that the overall performance of the Learning framework is impacted if one of the models in the framework does not perform as expected. In fact, stacking the models makes it possible to compensate for the shortcomings of each one while relying blindly on the qualities of each one.

Conclusion

Despite our unsatisfactory results, we believe that with more resources and a deeper understanding of the model architecture, an ensemble learning approach can improve the overall performance of meme classification.

It's worth remembering that it's not necessary to have an overly large model, even for this kind of complex classification. A well-trained model is preferable to a complex model.

References

[1] Phillip Lippe et al., "A Multimodal Framework for the Detection of Hateful Memes". 2020. arXiv:2012.12871 [cs.CL]
[2] Jason Brownlee. "A Gentle Introduction to Ensemble Learning Algorithms". In: MachineLearningMastery.com (2021). URL: https://machinelearningmastery.com/tour-of-ensemble-learning-algorithms/
[3] Hateful Memes Challenge and Dataset. https://ai.meta.com/tools/hatefulmemes/