Improving Hateful Memes Identification with Ensemble Learning

HENG Theo, SABATIER Aubin, VAN DEN HEUVEL Victor

***Abstract*—This document addresses the problem of multimodal classification by focusing on Ensemble Learning method. The aim is to improve the current performance of state of the art models by stacking multiple of them, each trained on a specific topic. The methods used to format the dataset, fine-tune the models and validate the results will be discussed. Finally, the limitations of this solution will be analysed.**

***Index Terms*—Deep Learning, Ensemble Learning, Stacking, Hateful Meme detection, Multimodal Model, CLIP, BLIP, ALIGN**

I. Introduction

The development of solutions for detecting hate content is an important part of creating a safer and more respectful environment. Hateful content on the Internet takes various forms, including audio, video, text, or images. This report focuses on developing a solution to handle memes, a type of media that conveys ideas or emotions through text overlaid on an image. Memes, being composed of two modalities, can express ideas implicitly or subtly, which poses a challenge in model development.

Defining offensive memes is a crucial step, especially during dataset annotation and model evaluation. Any memes propagating ideas or emotions aimed at discriminating against or harming the social identity of a community, gender, or social class are considered offensive. Defining this boundary can be very challenging, especially when dealing with jokes.

To train the various models, we have chosen to use the "Hateful Memes" dataset [6] provided by the Meta group. Our criteria for this choice were primarily annotation quality and the quality of the memes. The annotation process involved a third-party annotation company, where annotators received training and coaching to enhance their performance. The size of the dataset and the good distribution of memes between hateful and non-hateful categories were also determinant factors in this choice. Since the dataset is provided by Meta, we can easily compare our results with other approaches studied on the same dataset.

The solution explored in this report involves implementing an ensemble learning structure, specifically stacking, for hate content detection. Stacking entails training multiple models to detect hateful content for a specific topic. It is therefore possible to take into account datasets with very diverse content (like memes), even on a large scale, as each subject area will have its own specialized model. Moreover, it is possible to compensate for the weakness of each model.

The project's objective is to adapt the provided dataset, train models for each class, and compare the results against the same models fine-tuned for detecting hateful content across all topics. To ensure result quality, we will apply this process to three different model structures: CLIP, BLIP, and ALIGN.

II. Related Work

The challenge of detecting content across two types of data such as sound and videos or images and text, is not new. Therefore, leveraging previous research and understanding the limitations of existing approaches can lead to better results.

Several studies aim to develop multimodal models capable of handling information from multiple modalities, such as images and text, for hate content detection in memes. As explained in the article by [11], achieving better performance with multimodal models compared to models that process only text remains challenging. The goal is to enhance performance by implementing an ensemble learning structure.

The article [5] proposes an interesting approach, training different model types for images and text from the Memotion dataset. Although their results are not significant, they still outperform multimodal models trained on both images and text.

An evolutionary algorithm to optimize model weights implemented on the dataset from the group Meta allowed to improve the performances [9]. Despite being published in 2020, it remains a valuable benchmark. Additionally, the article [12] achieved third place on the Meta competition leaderboard using a similar ensemble learning method.

Unfortunately, comparing the results of these articles is challenging due to their use of different metrics to evaluate model performance, the choice of dataset, models, and ensemble learning structure which directly impacts the outcome. Techniques like bagging, stacking, or boosting are each suited to different applications.

The promising results from the presented articles highlight both the potential of using transformers in ensemble learning and the room for improvement. The objective is to apply a similar structure by training models on specific topic rather than specific data types.

III. Method

III-A Formatting the dataset

To be able to fine-tune the models on a particular class in the dataset, it was important to sort the dataset to group together the memes of the same class. Note that the text has already been extracted from the memes in the chosen dataset. It is therefore possible to sort them according to the keywords used in their description. Here are a few rule for the keyword selection:

- Use only common nouns (no determiners, linking words, etc.)
- Use only keywords that can target one type of population
- Do not use stereotypes or words with multiple connotations, as they may appear in unexpected subjects, or check them manually if necessary.

Even if it's impossible to take into account all the keywords, and in particular wordplay or abbreviations, it's important not to use keywords that may bias the training of the models. In addition, we need to keep track of meme identifiers so that they are not added several times to the same dataset. However, they can appear in several datasets. Subsequently, it was possible to assess which topics were most significant in the dataset. Given that data size plays a crucial role in achieving good performance during training, especially when dealing with diverse data, we decided to retain only topics with more than 500 memes. Consequently, our focus narrowed exclusively to detecting racist, misogynist, and Islamophobic content. Here are some examples of keywords for the three topics:

- Women: "ladies", "girls", "sisters", "wifes", "feminist", "mother", "women", "feminism", "sexism"
- Muslims: "muslims", "allah", "quran", "radical", "jihad", "terrorism", "arab", "islam", "islamic"
- Africans: "racists", "racist", "slaves", "color", "blacks", "africa", "african", "slavery"

Additionally, each of the created datasets consists of 50% of the targeted class and 50% of random memes from other topics.

We are also defining the "Base" topic. The "Base" train set results from merging the train sets of our three topics. This topic will serve as the training data for our BASE models, providing us with a performance baseline against which we will compare our Ensemble approach.

Below is a summary table of the composition of the four datasets used. The "harm rate" is defined as the ratio between the number of harmful memes related to the targeted class and the total number of memes in the dataset.

Topics	Size	Africans	Women	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 I: Datasets structure for fine tuning the models

Here we use 80% of each dataset to train the models and 20% to validate performance after each data epoch. Once the training is complete, we use the test set proposed by the dataset, containing never-before-seen memes, to test the classification.

Topics	Size	Africans	Women	Muslims	Harm rate
Test set A	142	60	0	0	24.65%
Test set W	275	0	54	0	6.93%
Test set M	186	0	0	61	13.98%
Test set Base	692	64	61	61	✗

Table II: Datasets structure for testing the models

We can see that the dataset made up of memes from the women's topic has a much lower harm rate than the other datasets, which can be a problem to train properly the models on these data. In order for the A, W and M models to recognise only one type of hate, all the labels in each dataset that does not concern the topic in question are set to zero, regardless of whether or not their content is hateful. Only the memes harmful and belonging to the topic are labeled as 1. The base dataset keeps its original label.

III-B Choice of the models

We chose to apply the stacking structure to 3 models with different architectures to ensure that the results obtained did not depend solely on the type of model. We have therefore chosen the following 3 state of the art multimodal models, available on the Hugging Face platform:

- **CLIP:** (Contrastive Language-Image Pre-Training) (2021) CLIP is trained on a variety of (image, text) pairs and employs a ViT-like transformer to extract visual features and a causal language model to obtain text features. These features are then projected into a latent space of the same dimension, and the similarity score is calculated using the dot product between the projected image and text features. [4] [10]
- **BLIP:** (Bootstrapping Language-Image Pre-training) (2022) BLIP is a framework designed for better vision-language understanding and generation, supporting a wider range of tasks than current methods. It features a Multimodal Mixture of Encoder-Decoder (MED) architecture, which can function as various types of encoders and decoders. Additionally, it introduces the Captioning and Filtering technique to improve training data quality by generating and filtering captions from noisy image-text pairs. [2] [8]
- **ALIGN:** (2021) ALIGN is a multi-modal vision and language model. It can be used for image-text similarity and for zero-shot image classification. ALIGN features a dual-encoder architecture with EfficientNet as its vision encoder and BERT as its text encoder, and learns to align visual and text representations with contrastive learning. Unlike previous work, ALIGN leverages a massive noisy dataset and shows that the scale of the corpus can be used to achieve SOTA representations with a simple recipe. [1] [7]

III-C Choice of strategy : Ensemble Learning

Stacking Ensemble Learning 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, ranging from 0 to 1 based on their confidence and the detected level of hate. The second part of the structure reconciles the estimates from each model by trusting the model with the higher estimation. Since each model has become specialized in a specific topic, they can only detect hateful memes related to their specific topic.

Contrary to the discussed papers, it is not possible to use a Majority Voting Classifier because the models are trained

[1] ALIGN, Hugging Face documentation, huggingface.co [1]

on a single topic, making them irrelevant for other topics. As a result, it becomes possible to detect both hateful content and targeted populations.

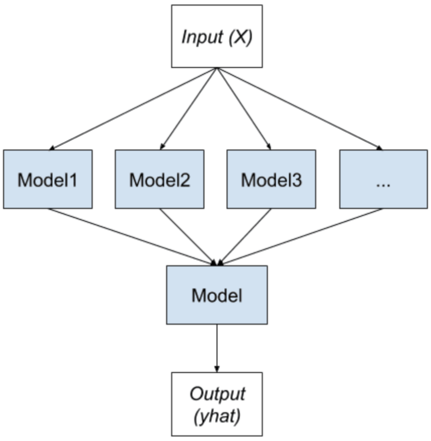

Figure 1: Stacking Ensemble Learning architecture [3]

To test our hypothesis, we therefore need to train three models on each of the four dataset created, making a total of 12 models.

III-D Training the models

Fine-tuning Given the size and complexity of the models chosen, we opted for model fine-tuning. At the test stage, we first attempted to fine-tune only the classification head while freezing the rest of the network. This approach gave poor results. Later, for the BLIP model, layers 7 to 11 of the vision encoder and text encoder were frozen, while for ALIGN we froze the last 9 layers (45 to 54) of the vision encoder and 7 to 11 of the text encoders. CLIP, for its part, has been fine-tuned across all its layers. We kept these choice as a trade-off between the best possible performance and computing capacity.

Learning rate To improve the training process, we chose a learning rate modulated by a scheduler. Depending on the variation in loss after each epoch, a high rate causes sudden jumps in model training and compromises convergence. The scheduler can be used to reduce the rate when the loss is low.

Batch size Using a large batch size means that the model can be trained on more images and a more stable, less noise-sensitive gradient can be obtained. We opted for a batch size of 32, which gives us a good balance between performance and computing cost. We have repeatedly encountered memory problems caused by larger batch sizes and have therefore not managed to increase them significantly.

Hyperparameters	CLIP	BLIP	ALIGN
Learning rate	5e-5	5e-6	5e-5
Batch size	32	32	32

Table III: Choice of hyperparameters for training

III-E Testing the results

The models are tested using the F1-score using unseen test data with the original labeling.

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 IV: F1-score of the models on unseen Base test data.

We can see that $model_A$, $model_W$ and $model_M$ 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. Unfortunately, it was not possible to train the ALIGN model on the dataset W and the dataset Base.

III-F Analysis

For a model such as CLIP, the use of ensemble learning improved the F1 score by 2.25%. For BLIP, however, the score with ensemble learning is noticeably lower. As such, we cannot really conclude on the effectiveness of ensemble learning from our study.

Training and testing depend notably on the random initialization of the training and testing sets (80/20%). The effect is similar when shuffling before creating the batches. We noticed significant changes in performance on different runs.
This may suggest that there are too few memes to train well, or that there is too much intra-class variation, which might prevent the model from being fine tuned correctly with too few different memes. These changes in performance also appear to be linked to the topics: the performance obtained for models linked to the topic "Women", for example, is generally lower than that for the topic "Muslims".

Furthermore, training appeared to be particularly difficult for certain models for which the resources required were substantial. The ALIGN model, for example, did not allow us to run the training for all datasets because of memory overflow (even on an A100 GPU with 40GB of RAM).

Finally it should be noted that if one model has weak performance, it can have a negative impact on the overall performance. As said this is the case for the "Women" set that penalizes the overall results.

IV. Conclusion

During this project we have seen how important the size and the quality of the dataset is in order to train properly a neural network. 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. We also faced the impact of the batch size, learning rate and model architecture during our research, teaching us how to make compromises.

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] *ALIGN*. https://huggingface.co/docs/transformers/model_doc/align.

[2] *BLIP*. https://huggingface.co/docs/transformers/model_doc/blip.

[3] Jason Brownlee. "A Gentle Introduction to Ensemble Learning Algorithms". In: *MachineLearningMastery.com* (2021). URL: https://machinelearningmastery.com/tour-of-ensemble-learning-algorithms/.

[4] *CLIP*. https://huggingface.co/docs/transformers/model_doc/clip.

[5] Amitava Das, Amit Sheth, and Asif Ekbal. "De-Factify 2: 2nd Workshop on Multimodal Fact Checking and Hate Speech Detection". In: *Proceedings of De-Factify 2: 2nd Workshop on Multimodal Fact Checking and Hate Speech Detection, co-located with AAAI 2023*. 2023. URL: https://ceur-ws.org/Vol-3555/paper6.pdf.

[6] *Hateful Memes Challenge and Dataset*. https://ai.meta.com/tools/hatefulmemes/.

[7] Chao Jia et al. *Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision*. 2021. arXiv: 2102.05918 [cs.CV].

[8] Junnan Li et al. *BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation*. 2022. arXiv: 2201.12086 [cs.CV].

[9] Phillip Lippe et al. *A Multimodal Framework for the Detection of Hateful Memes*. 2020. arXiv: 2012.12871 [cs.CL].

[10] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: 2103.00020 [cs.CV].

[11] Shardul Suryawanshi et al. "Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text". In: *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying* (2020), pp. 32–41. URL: https://aclanthology.org/2020.trac-1.6.pdf.

[12] Riza Velioglu and Jewgeni Rose. *Detecting Hate Speech in Memes Using Multimodal Deep Learning Approaches: Prize-winning solution to Hateful Memes Challenge*. 2020. arXiv: 2012.12975 [cs.AI].