Machine Learning for Humanities – project report

Academic year 2024/2025

Niccolò Eros Molinati

1. Introduction and benchmark

The project aims at tackling the Hateful Memes Challenge launched by META in 2020 and reported in the paper by D. Kiela et al[1]. The challenge consists in creating and optimizing a model apt to recognize and label instances of hate speech online, specifically in the form of *memes*. By memes, we refer to a multimodal virtual object composed of an image and an overlaid text (or caption), whose parts collaborate in creating a meaning that can't be reduced simply to their sum. Due to the high inferential effort required to correctly analyze a meme and to the ambiguity of the message conveyed, the task isn't easy to approach.

The aforementioned paper shows different methods in facing the challenge, that here will be quickly reminded: In addition to the monomodal approach (which uses only a text or image model to make predictions), there are multimodal approaches with monomodal pretraining and multimodal pretraining.

For the project, a model inspired by the multimodal with monomodal pretraining approach (and more specifically, from Late Fusion and Concat BERT) was implemented, although it does not perform as optimally as one leveraging multimodal pretraining, in order to leave enough space for the implementation aspect of this exam.

2. The Dataset

The dataset was manually created by annotators for the purpose of the challenge. However, it is no longer available on the dedicated website, as access was limited to those who participated in the challenge in 2020. Fortunately, versions of the original dataset were uploaded to different platforms, and the one used for this project was available on Hugging Face.[2].

However, there are some critical points to address regarding its quality. First of all, it appears clear that the dataset on Hugging Face is only a fraction of the one designed by META, as it contains roughly only eight thousand instances (against the ten of the original one). Furthermore, upon closer inspection, it appears that not all instances in the JSON-LD file reference an image in the dedicated folder, and vice versa. Data cleaning operations revealed that the dataset was reduced to just 6,744 instances, leading to various problems that were only marginally mitigated. Given the complexity of the task and the architecture of my model, which requires multiple steps of fine-tuning and training (more on that in the next section), the dataset's scarcity could lead to overfitting. To avoid it, limiting the number of training epochs and applying *K-fold cross-validation*. Cross-validation, used in training my model, is a well-established method for handling datasets with insufficient data[3] and was a crucial step in improving the model.

Another significant result of the aforementioned issue is the imbalance of the dataset: the paper by D. Kiela et al. clearly illustrates how the original dataset was crafted:

> «40% multimodal hate, 10% unimodal hate, 20% benign text confounder, 20% benign image confounder, 10% random non-hateful.»[4]

These proportions are not respected in the version at hand. A simple query reveals that approximately 30% of the data is labeled hateful (about 1,950 instances), compared to the original 50%. No real measure was taken

[1]Douwe Kiela et al., *The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes*, 2021.
[2] Available at the following link: https://huggingface.co/datasets/neuralcatcher/hateful_memes.
[3] Zhang et al., *Dive Into Deep Learning*, MIT Press, 2023. See section 3.6.3.1.
[4] Douwe Kiela et al., *The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes*, 2021, p. 5.

to address this issue, as the missing data could not be found even after consulting different versions on other platforms. What we can do, however, is contextualize the results.

Before moving to the model and its structure, some space will be left for a brief introduction of the dataset. All memes were hand-crafted and consist of both a textual caption and an image. Some can be considered hateful due to the meaning 'created' by the union of both elements (multimodal hate), while others are labeled hateful based on only one of the two components. Since, in many cases, the 'hatefulness' of a meme is due to its overall composition, some benign confounders were added. Either the text or the image may appear in a hateful instance, but in this context, they take on a different meaning, and the meme ends up being non-hateful.

When discussing what 'hateful' means in the context of a meme, there is no easy answer. The original paper provides an operational definition to be used as a criterion during the construction of the dataset:

> « A direct or indirect attack on people based on characteristics, including ethnicity, race, nationality, immigration status, religion, caste, sex, gender identity, sexual orientation, and disability or disease. We define attack as violent or dehumanizing (comparing people to non-human things, e.g. animals) speech, statements of inferiority, and calls for exclusion or segregation. Mocking hate crime is also considered hate speech.»

While it doesn't completely resolve the ambiguity of the concept, it provides a concrete reference for understanding and evaluating the predictions of our model.

3. Workflow and architecture

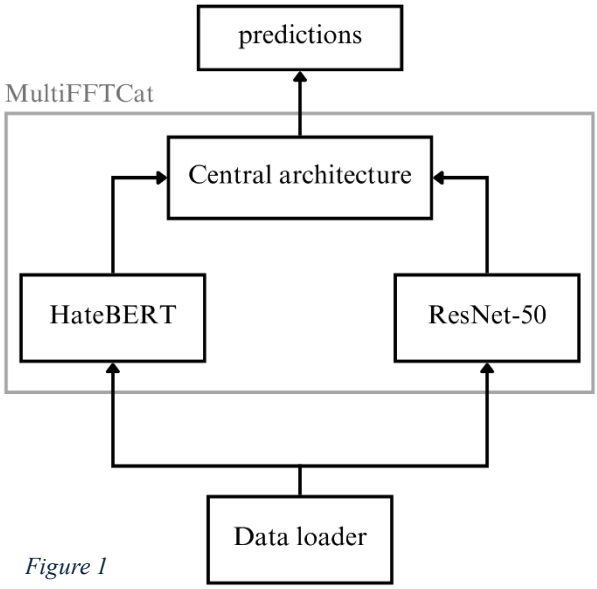

Figure 1

The model presented in this report, which will from now on be referred to as MultiFFTCat, is inspired by CatBERT in the context of monomodal pretraining. It incorporates a ResNet-50 and a HateBERT, whose predictions are then processed and fused in the "central" part of the architecture (fig.1).

Both pretrained models are finetuned on their classifier layer (adding a fully connected layer for ResNet) with 25% of the dataset. The following operations are then performed on the forward:

1. Obtaining predictions from ResNet and HateBERT.
2. Projection of said predictions (kept separated).
3. Performing FFT on the projections.
4. Fusion via concatenation.
5. Final transition through MLP for predictions.

Between each linear layer dropout and ReLu operations are performed to strengthen the model and mitigate overfitting[5]. The intuition here is that projecting logits from pretrained parts before concatenation will result in a more complex rendition of the interaction between them, and the model can be trained to weigh those representations.

Leveraging the Fourier Transform to augment generalization capabilities, especially in image data, has been richly documented[6]; the scope of applying it after the projection is to perform further abstraction in order to capture significant patterns and correlations between the two modalities. During model selection through

[5] Srivastava et al., *"Dropout: A Simple Way to Prevent Neural Networks from Overfitting"* (JMLR 2014).
[6] Qinwei Xu et al., *A fourier-based framework for domain generalization*, 2021; Hongyi Pan et al., *Domain Generalization with Fourier Transform and Soft Thresholding*, 2023; Yusuke Iwasawa et al., *Deep Frequency Filtering for Domain Generalization*, 2023; Varsha Nair et al., *Fast Fourier Transformation for Optimizing Convolutional Neural Networks in Object Recognition*, 2020.

testing, this proved to be an architecture resulting in fairly consistent improvement, compared to different orders of the same operations or exclusion of FFT (especially after the introduction of gradient clipping[7]).

The workflow is divided in the following steps: 1) fine-tuning HateBERT and ResNet (25% of dataset), 2) training with K-fold cross-validation (68%) and 3) testing (remaining 7%). Other methods were implemented to manage the issues concerning scarcity of data – specifically, weight decay[8] and class weighting in the loss.

4. Results and conclusion

In the paper introducing the challenge, the average results for multimodal with monomodal pretraining are reported in table 1 and compared with those of MultiFFTCat. However, we could argue that the imbalance of the dataset used for the project requires us to carefully contextualize our findings.

Model	**Val acc.**	**Val roc-auc**	**Test acc.**	**Test roc-auc**
Late Fusion	59.39	65.07	63.20±1.09	69.30±0.33
Concat BERT	59.32	88	61.53±0.96	67.77±0.87
MultiFFTCat	67.57±1.0	68±3	69.37	71 (fig.2)

Table 1

Considering the proportions in which labels appear, we see that, we understand that MultiFFTCat is performing significantly worse than the Late Fusion and Concat BERT (if the model guessed only 0, it would reach almost 66%); however, learning is taking place: each cycle of training results in a small yet consistent improvement. Even more, during testing, increasing the hyperparameter for the number of folds (and thus the cycles) results in a similar improvement. In conclusion, it is worth noting that measures previously cited (K-fold cross-validation, weight decay, weighted loss, gradient clipping) all resulted in noticeable improvement during all stages of training and validation.

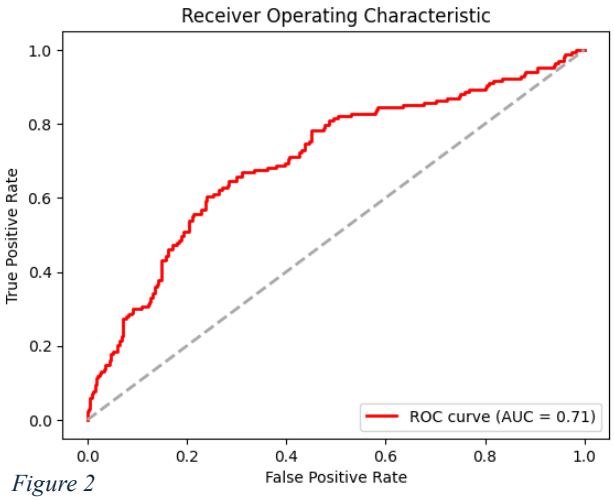

Figure 2

To conclude, the project includes an example of application of the model, in which a simple function extracts predictions and decides whether to alert the inclusion of hateful content in the image. Given the challenges in determine the concept of "hate speech" (despite the definition given at the start of this report) it is implemented to censor an image over a confidence threshold.

5. Bibliography

- Neuralcatcher (2020), *hateful_memes* [Dataset]. Hugging Face. Link.

[7] Zhang et al., *Dive Into Deep Learning*, MIT Press, 2023. See section 9.5.3.
[8] A. Lo, *Weight Decay and Its Peculiar Effects*, article on Medium, Dec. 18, 2021.

- Douwe Kiela et al.(2021), *The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes*, arXiv preprint arXiv: 2005.04790.https://arxiv.org/abs/2005.04790.
- Zhang et al., *Dive Into Deep Learning*, MIT Press, 2023.
- Srivastava et al., *"Dropout: A Simple Way to Prevent Neural Networks from Overfitting"*, JMLR, 2014.
- Qinwei Xu et al. (2021), *A fourier-based framework for domain generalization*, arXiv preprint arXiv: 2105.11120. https://arxiv.org/abs/2105.11120.
- Hongyi Pan et al. (2023), *Domain Generalization with Fourier Transform and Soft Thresholding*, arXiv preprint arXiv: 2309.09866. https://arxiv.org/abs/2309.09866.
- Yusuke Iwasawa et al. (2023), *Deep Frequency Filtering for Domain Generalization*, arXiv preprint arXiv: 2203.12198. https://arxiv.org/abs/2203.12198.
- Varsha Nair et al. (2020), *Fast Fourier Transformation for Optimizing Convolutional Neural Networks in Object Recognition*, arXiv preprint arXiv: 2010.04257. https://arxiv.org/abs/2010.04257.
- A. Lo, *Weight Decay and Its Peculiar Effects*, Towards Data Science, Dec 10, 2021.