Multimodal or Text? Retrieval or BERT? Benchmarking Classifiers for the Shared Task on Hateful Memes

Vasiliki Kougia
Stockholm University
vasiliki.kougia@dsv.su.se

John Pavlopoulos
Stockholm University
ioannis@dsv.su.se

Abstract

The Shared Task on Hateful Memes is a challenge that aims at the detection of hateful content in memes by inviting the implementation of systems that understand memes, potentially by combining image and textual information. The challenge consists of three detection tasks: hate, protected category and attack type. The first is a binary classification task, while the other two are multi-label classification tasks. Our participation included a text-based BERT baseline (TxtBERT), the same but adding information from the image (ImgBERT), and neural retrieval approaches. We also experimented with retrieval augmented classification models. We found that an ensemble of TxtBERT and ImgBERT achieves the best performance in terms of ROC AUC score in two out of the three tasks on our development set.

1 Introduction

Multimodal classification is an important research topic that attracts a lot of interest, especially when combining image and text (Li et al., 2019; Lu et al., 2019; Chen et al., 2020; Gan et al., 2020; Su et al., 2019; Yu et al., 2020; Li et al., 2020). Humans understand the world and make decisions, by using many different sources. Hence, it is reasonable to infer that Artificial Intelligence (AI) methods can also benefit by combining different types of data as their input (Gomez et al., 2020; Vijayaraghavan et al., 2019). The Hateful Memes Challenge and dataset were first introduced by Facebook AI in 2020 (Kiela et al., 2020). The goal was to assess multimodal (image and text) hate detection models. The dataset was created in a way such that models operating only on the text or only on the image would not have a good performance, giving focus to multimodality (see Section 2). The winning system used an ensemble of different vision and language transformer models, which was further enhanced with information from input objects detected in the image and their labels (Zhu, 2020). The Hateful Memes shared task extends this competition by adding fine-grained labels for two multi-label tasks (see Fig. 1). The first task is to predict the protected category and the second to predict the attack type.

Figure 1: An example of a hateful (left) and a not hateful (right) meme. ©Getty Images

2 Dataset

The provided dataset comprises images and text. First, Kiela et al. (2020) collected real memes from social media, which they called source set and then, used them to create new memes. For each meme in the source set, the annotators searched for images that had similar semantic context with the image of the meme and replaced the image of the meme with the retrieved images.[1] The newly developed memes were then annotated as hateful or not by the annotators. For the hateful memes, counterfactual examples were created and added to the dataset

[1]The similar images come from Getty Images (https://www.gettyimages.com/).

Proceedings of the Fifth Workshop on Online Abuse and Harms, pages 220–225

by replacing the image or the text. Following this process a dataset of 10,000 memes was created.

For the Shared Task on Hateful Memes at WOAH 2021, the same dataset was used, but with additional labels. New fine-grained labels were created for two categories: protected category and attack type. Protected category indicates the group of people that is attacked in a hateful meme and consists of five labels: race, disability, religion, nationality and sex. The attack type refers to the way that hate is expressed and consists of seven labels: contempt, mocking, inferiority, slurs, exclusion, dehumanizing, inciting violence . If a meme is not hateful, then the pc_empty label is assigned for the protected category task and the attack_empty label for the attack type task. A meme can have one or more labels, leading to a multi-label classification setting.

Participants of the shared task were provided with a training set comprising 8,500 image-text pairs and two development datasets with 500 and 540 image-text pairs. In our work, we merged these sets and split the total of 9,140 unique pairs to 80% for training, 10 % for validation and 10 % as a development set. The unseen test set for which we submitted our models' predictions consisted of 1,000 examples. The dataset was imbalanced, with approximately 64% of the memes being not hateful.

3 Methods

The methods we implemented for this challenge comprise image and text retrieval, BERT-based text (and image) and retrieval-augmented classification (RAC). The following subsections describe the implemented methods.

3.1 Retrieval

Multimodal Nearest Neighbour (MNN) employs image and text retrieval. In specific, for an unseen test meme, MNN retrieves the most similar instance from a knowledge base (here, the training dataset) and assigns its labels to the unseen meme.

We used two MNN variants, which differed in the way they encode the text. For the encoding of images, each variant used a DenseNet-121 Convolutional Neural Network (CNN), pre-trained on ImageNet (Deng et al., 2009). Each CNN was fine-tuned for the corresponding task independently on our data. For the encoding of text, the first variant uses the centroid of Fasttext word embeddings for English pre-trained on Common Crawl (Grave et al., 2018) (MNN:base).[2] The second variant employs three BERT models, each fine-tuned on one of our tasks (see subsection 3.2), from which we extracted the CLS tokens as the representation of memes' texts (MNN:BERT).

The similarity between the query embeddings (both, image and text) and the knowledge base is computed using the cosine similarity function. During inference, given a test meme, we find the most similar training image to the meme image and the most similar training text to the meme text. Then, we retrieve the labels of these two retrieved training examples. If a label appears in both examples, it is assigned a probability of 1. If it appears in only one example it is assigned the cosine similarity of that example. The rest of the labels, are assigned a zero probability.

3.2 BERT-based

For this method we also tried two text and one multimodal approach. The first text-based approach (TxtBERT) takes as input only the text of the meme. The second, dubbed CaptionBERT, takes as input the meme text and the image caption, separated with the [SEP] pseudo token. We employed BERT base for both and fine-tuned it on our data (one for each task). The image captions were generated by the Show and Tell model (S&T) (Vinyals et al., 2015), which was trained on MS COCO (Lin et al., 2014). In both approaches we extract the [CLS] pseudo-token and feed it to a linear layer that acts as our classifier.

The multimodal approach (ImgBERT) combines TxtBERT above with image embeddings, which are extracted by the same CNN encoder that was used for MNN (see subsection 3.1). We concatenate each image embedding with the BERT representation of the [CLS] pseudo token and feed the resulting vector to the classifier.

The outputs of the classifier correspond to the labels for the multilabel classification tasks and each output is passed through a sigmoid function, in order to obtain one probability for each label. In the binary classification task the output is one probability, where 1 means the text is hateful and 0 means it is not. The BERT-based models are trained using binary cross entropy loss and the Adam optimizer with learning rate 2e-5. Early stopping is applied

[2] https://fasttext.cc/docs/en/crawl-vectors.html

during training with patience of three epochs.

3.3 RAC-based

Inspired by retrieval-augmented generation (RAG) (Lewis et al., 2020), we experimented with Retrieval Augmented Classification (RAC), in order to expand the knowledge of our BERT-based models and improve their performance. To do that we combined TxtBERT and ImgBERT with MNN retrieval and call the two new methods TxtRAC and Txt+Img RAC respectively. The most similar text obtained by MNN:BERT is concatenated to the text of the meme, separated with the [SEP] pseudo-token, and it is passed to TxtBERT (in TxtRAC) and ImgBERT (in Txt+Img RAC). The training setup is the same as the one in the BERT-based models described above (see Section 3.2).

3.4 Ensemble

An ensemble was created combining visual and textual information, based on ImgBERT and TxtBERT. For each label of each task, the ensemble averages the two scores, one per system.

4 Experimental Results

The official evaluation measure of the shared task is the ROC AUC score. Hence, we provided the output probability distribution over the labels of each task from a model in order to evaluate it. The classifiers of our models did not output a probability for the corresponding empty label (meaning that the meme is not hateful) of each task. In order to assign a probability to the not hateful label of the binary classification task we compute 1 - hateful probability. To the pc_empty and attack_empty labels of the corresponding task, we assign the probability of 1 - maximum probability of the other labels. The provided evaluation script computes the ROC AUC score micro averaged and with the one-vs-rest method. It also computes the micro F1 score by applying a threshold (0.5) to the predicted probabilities.

Each team participating in the Shared Task on Hateful Memes could submit predictions from two systems on the unseen test set. We chose to submit the TxtBERT and the ensemble of TxtBERT and ImgBERT.[3] In Table 4 we present the results on the hidden test set. The organizers provided us the ROC AUC scores for the protected category and the attack type tasks. Since we do not have the gold labels of the test set in order to evaluate all the models we implemented, we report their results on the development set we created. Table 1 presents the evaluation scores for the hate task on our development set, Table 3 for the attack type task, and Table 1 for the protected category task. Moreover, in Tables 5 and 6 we report the F1 and ROC AUC scores for each label of the protected category and attack type tasks respectively.

[3]The code for our two submitted models is available at: https://github.com/vasilikikou/hateful_memes

Model	F1	AUC
TxtBERT	0.755	0.821
CaptionBERT	0.724	0.780
MNN:base	0.674	0.617
MNN:BERT	0.704	0.663
ImgBERT	0.689	0.755
TxtRAC	0.702	0.799
Txt+Img RAC	0.712	0.796
Ensemble	**0.765**	**0.863**

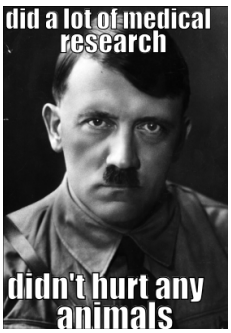
Table 1: Micro F1 and ROC AUC scores of our models for the binary classification "hateful or not" task. In this task the ensemble of TxtBERT and ImgBERT outperforms all other methods.

Model	F1	AUC
TxtBERT	**0.729**	**0.931**
CaptionBERT	0.724	0.920
MNN:base	0.566	0.783
MNN:BERT	0.578	0.794
ImgBERT	0.640	0.818
TxtRAC	0.717	0.927
Txt+Img RAC	0.640	0.840
Ensemble	0.694	0.920

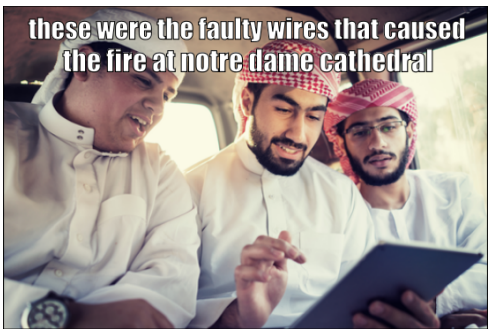
Table 2: Micro F1 and ROC AUC scores of our models for the protected category task. TxtBERT is the best performing model in this task.

Model	F1	AUC
TxtBERT	**0.681**	0.929
CaptionBERT	0.656	0.914
MNN:base	0.559	0.798
MNN:BERT	0.600	0.825
ImgBERT	0.666	0.928
TxtRAC	0.665	0.925
Txt+Img RAC	0.662	0.928
Ensemble	0.670	**0.932**

Table 3: Micro F1 and ROC AUC scores of our models for the attack type task. The ensemble achieves the best AUC and TxtBERT the best F1 score.



(a) 'hateful', 'religion', 'mocking'



(b) 'hateful', 'religion', 'dehumanizing'

(c) 'hateful', 'religion;nationality', 'exclusion'

(d) 'hateful', 'nationality', 'dehumanizing

Figure 2: The two memes on top (a, b) were better classified by ImgBERT while the two memes below (c, d) by TxtBERT. Ground truth in captions. ©Getty Images

Model	Protected category	Attack type
TxtBERT	**0.876**	0.881
Ensemble	0.865	**0.890**

Table 4: ROC AUC scores of our two submissions for the protected category and attack type tasks as provided by the organizers.

5 Discussion

MNN:BERT outperforms MNN:base in all three tasks. This is probably due to the fact that a simple centroid of word embedding ignores word order, by contrast to a BERT-based representation, which also encodes the position of the word. Interestingly, CaptionBERT outperformed ImgBERT both in hate and protected category detection. This means that integrating the automatically generated caption of the image, instead of the image itself, was beneficial for two out of three tasks. In attack type detection, however, this didn't apply. We also observe that employing the most similar text in the TxtBERT model (TxtRAC), leads to a worse performance, showing that the retrieved text does not help the text classification model as expected. This probably occurs due to the diversity of the texts in the dataset. However, TxtRAC outperforms CaptionBERT in all tasks in terms of ROC AUC, maybe because generated captions from S&T, which is only trained on MS COCO can contain errors.

The ensemble model, that averages the predictions of TxtBERT and ImgBERT, outperformed the rest of the models, in ROC AUC, for hate and attack type detection. However, we note that for a fair comparison we should have created also checkpoint-based ensembles per model. That is, we can't be certain whether the superior performance of the ensemble stems from the combination of textual and visual information or from the reduction of the variance of the models that are used by the ensemble.

In the ROC AUC scores for the hidden test set (see Table 4), we observe similar performance of the models as in the development set. In particular, TxtBERT achieves the best score for the protected category task, while the Ensemble is the best for the the attack type task.

For the two multilabel tasks we also evaluated our models per label in order to obtain a better understanding of their performance. We observe that even though the dataset is imbalanced containing more not hateful memes, the scores of the models for the empty label are lower than the ones for the other labels in both tasks. This means that the models do not achieve a very high performance on the empty label as expected. Also, we see that there

Model	empty		religion		sex		race		disability		nationality	
	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC
TxtBERT	0.808	0.776	0.609	**0.875**	**0.663**	**0.913**	**0.595**	0.873	**0.400**	**0.843**	**0.351**	0.912
CaptionBERT	**0.824**	0.746	0.406	0.869	0.634	0.909	0.479	0.854	0.158	0.765	0.061	0.895
MNN:base	0.767	0.530	0.354	0.678	0.313	0.649	0.224	0.566	0.244	0.635	0.138	0.564
MNN:BERT	0.787	0.590	0.348	0.663	0.282	0.624	0.234	0.574	0.217	0.608	0.096	0.536
ImgBERT	0.789	0.414	0.000	0.661	0.000	0.609	0.000	0.385	0.000	0.632	0.000	0.544
TxtRAC	0.803	**0.794**	**0.631**	0.871	0.630	0.907	0.610	**0.879**	0.000	0.773	0.154	0.859
Txt+Img RAC	0.789	0.606	0.000	0.633	0.000	0.670	0.000	0.573	0.000	0.723	0.000	0.573
Ensemble	0.821	0.759	0.107	0.858	0.422	0.890	0.380	0.838	0.000	0.837	0.000	**0.927**

Table 5: F1 and ROC AUC scores per label for the protected category task. There are five labels for this task and the empty label for not hateful memes.

Model	empty		mock.		deh.		viol.		cont.		excl.		inf.		slurs	
	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC
TxtBERT	**0.811**	**0.778**	**0.491**	0.870	0.416	0.814	0.000	0.815	0.000	0.926	0.000	0.678	**0.354**	0.756	0.829	**0.986**
CaptionBERT	0.804	0.708	0.449	**0.883**	0.242	0.779	0.000	0.774	0.000	0.848	0.000	0.645	0.200	0.695	0.087	0.949
MNN:base	0.770	0.521	0.303	0.703	0.295	0.578	0.172	0.577	0.237	0.661	0.111	0.552	0.265	0.628	0.277	0.695
MNN:BERT	0.793	0.571	0.326	0.707	0.332	0.600	**0.220**	0.609	**0.351**	0.741	**0.114**	0.552	0.283	0.635	0.557	0.886
ImgBERT	0.791	0.750	0.444	0.841	0.427	0.803	0.207	0.852	0.000	0.931	0.000	**0.724**	0.350	0.747	0.837	0.971
TxtRAC	0.797	0.775	0.444	0.873	0.403	0.799	0.000	0.816	0.000	0.817	0.000	0.661	0.148	0.751	0.821	0.972
Txt+Img RAC	0.795	0.773	0.440	0.859	**0.457**	0.813	0.000	**0.858**	0.000	0.841	0.000	0.675	0.304	**0.760**	0.829	0.984
Ensemble	0.801	0.774	0.436	0.863	0.398	**0.820**	0.115	0.841	0.000	**0.933**	0.000	0.704	0.336	0.756	**0.857**	0.980

Table 6: F1 and ROC AUC scores per label for the attack type task. The labels for this task are seven: mocking (mock.), dehumanizing (deh.), inciting_violence (viol.), contempt (cont.), exlusion (excl.), inferiority (inf.), slurs and the emply label.

is not a clear winner, since for each label different models can have the best score. Besides TxtBERT and Ensemble, which have the best performance in the micro averaging setting, we see that other models can be better on specific labels. In particular, in the protected category task TxtRAC achieves the best ROC AUC score for the empty and race labels, showing that RAC can benefit these two categories. Interestingly, in the attack type task, retrieval also works well for the inciting_violence and inferiority labels, where Txt+Img RAC has the best ROC AUC score. CaptionBERT and ImgBERT have the best scores for the mocking label and the eclusion label respectively.

Error analysis

TxtBERT outperforms ImgBERT in all three tasks. In order to explain this observation in a meaningful way we compare the ROC AUC scores of several cases from the development set and see in which the image helped the classifier. We studied this for the hateful memes in our development set and saw that ImgBERT outperformed TxtBERT in only 8% of these memes. In Figure 2 we see two memes that ImgBERT predicted with a score closer to the ground truth than TxtBERT (above) and two memes that TxtBERT was closer to the ground truth (below). Indeed for the top two memes (a, b) we observe that the text on its own is not hateful, but when combined with the image a hateful meme is resulted. The third meme (c) has a text that contain slurs, which probably makes it easier for BERT to predict that it is hateful, while the image on its own is not. In the fourth meme (d), it is not clear that the text is hateful, but still TxtBERT is better in detecting this.

6 Conclusions

We participated in the Shared Task on Hateful Memes with the aim of detecting memes with hateful content, as well as the protected categories and the attack types in hateful memes. We experimented with models that employ only the text, that employ the text and image, and with models that also add information from retrieved texts. TxtBERT, a BERT for sequence classification that uses only the text, achieves very good performance. An ensemble of TxtBERT and a multimodal BERT (ImgBERT) outperforms all other methods on our development set in two out of the three tasks. We found that retrieval methods based on both the image and the text do not work well on this dataset, probably due to its complex context and diversity. In future work we plan to experiment with large pre-trained vision and language transformer models, different sources for retrieval and explainability approaches for multimodal methods.

References

Y-C Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu. 2020. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120, held on-line.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, Miami Beach, FL, USA.

Z. Gan, Y-C Chen, L. Li, C. Zhu, Y. Cheng, and J. Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. *arXiv:2006.06195*.

R. Gomez, J. Gibert, L. Gomez, and D. Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1470–1478, Aspen, CO, USA.

E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3483–3487, Miyazaki, Japan.

D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *34th Conference on Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada.

P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W-t Yih, T. Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv:2005.11401*.

L. H. Li, M. Yatskar, D. Yin, C-J Hsieh, and K-W Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv:1908.03557*.

X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137, held on-line.

T-Y Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C L Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755, Zurich, Switzerland.

J. Lu, D. Batra, D. Parikh, and S. Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv:1908.02265*.

W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv:1908.08530*.

P. Vijayaraghavan, H. Larochelle, and D. Roy. 2019. Interpretable multi-modal hate speech detection. In *International Conference on Machine Learning, AI for Social Good Workshop*, Long Beach, CA, USA.

O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. 2015. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, Boston, MA, USA.

F. Yu, J. Tang, W. Yin, Y. Sun, H. Tian, H. Wu, and H. Wang. 2020. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *arXiv:2006.16934*.

R. Zhu. 2020. Enhance multimodal transformer with external label and in-domain pretrain: hateful meme challenge winning solution. *arXiv:2012.08290*.