RIT Boston at SemEval-2022 Task 5: Multimedia Misogyny Detection By Using Coherent Visual and Language Features from CLIP Model and Data-centric AI Principle

Lei Chen*
Rakuten Institute of Technology (RIT)
Boston, MA
lei.a.chen@rakuten.com

Houwei Chou*
RIT
houwei.chou@rakuten.com

Abstract

Detecting MEME images to be misogynous or not is an application useful on curbing online hateful information against women. In the SemEval-2022 Multimedia Automatic Misogyny Identification (MAMI) challenge, we designed a system using two simple but effective principles. First, we leverage on recently emerging Transformer models pre-trained (mostly in a self-supervised learning way) on massive data sets to obtain very effective visual (V) and language (L) features. In particular, we used the CLIP (Radford et al., 2021) model provided by OpenAI to obtain coherent V and L features and then simply used a logistic regression model to make binary predictions. Second, we emphasized more on data rather than tweaking models by following the data-centric AI principle. These principles were proven to be useful and our final macro-F1 is 0.778 for the MAMI task A and ranked the third place among participant teams.

1 Introduction

Systematic inequality and discrimination to women does not appear offline but also in online communication. MEME is an image characterized by a visual content with an overlaying text added MEME creators. Although most of MEMEs are created with the intention of making funny jokes, some of MEMEs are created as a form against women. Therefore, identifying misogynous MEMEs is important for curbing such misuse.

In the SemEval-2022, the task 5, Multimedia Automatic Misogyny Identification (MAMI) (Fersini et al., 2022), was organized for this purpose. The challenge consists of two sub-tasks, including the task A, which is determining a MEME be misogynous or not, and the task B which distinguishes non-misogynous and 4 overlapped misogyny subtypes.

We participated in the MAMI challenge and focused on the task A. Our solutions focused on exploring various pre-trained Transformer models for extracting textual and visual features and utilizing a simple logistic regression (LR) model to make binary predictions. In addition, following a new trend in AI research, which is relying on the power provided by data more, i.e., data-centric AI (Ng, 2021b,a), we expanded available training data by manually marking more samples from the evaluation set. As a result, by jointly utilizing these methods, our team ended up on obtaining the third rank in the task A.

2 Related work

Automatic Misogyny Identification (AMI) has become an active research topic in natural language processing (NLP). For example, in the IberEval-2018, the AMI task was introduced as a new task (Fersini et al., 2018). The task consists of the three sub-tasks, i.e., misogyny identification, misogynistic behavior categorization, and target classification. The misogyny related annotations are made on both Spanish and English tweets. Among 11 different teams from 5 countries, (Pamungkas et al., 2018) ranked first in the misogyny identification task on both languages. It proposed an SVM-based architecture and explored several sets of features, including lexical features relying on the lexicon of abusive words.

Another AMI challenge (Fersini et al., 2020) was organized at the Evalita-2020 evaluation campaign and used Italian tweets. Its sub-task A is about misogyny and aggressiveness identification (4 class labels). A total 8 teams from 6 different countries participated in the challenge. Though a few teams used traditional word embedding to be textual features, most of the participants explored richer sentence embedding such as BERT (Devlin et al., 2019) or Roberta. Regarding modeling, the used methods are diverse, ranging from simple lo-

*Equal contributor

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 636 - 641

gistic regression (LR), to Convolutional Neural Network (CNN), even to fine-tuning pre-trained models.

Data plays an important role on AMI research development. How to organize misogyny labels is an important research question. (Guest et al., 2021) created a new dataset for tackling online misogyny. Its dataset consists of $6,567$ labels for Reddit posts and comments. A new hierarchical taxonomy has been proposed and a careful training was provided to annotators for obtain high-quality labels.

In the misogyny detection works described above, only textual clues are utilized. In a related topic, detecting hateful speech, image clues have been widely used to better reflect the fact that human communication is naturally multimodal. A Hateful Memes Challenge competition was held at NeurIPS 2020 (Kiela et al., 2020). The task is to classify a meme (i.e., an image and associated texts) to be hateful or not. In the challenge, a set of language-visual pre-trained models, such as UNITER, VILLA, and ERNIE-ViL, have been widely used for extracting semantically coupled textual and visual features. For example, the winning solution utilized four types of VL transformers (Zhu, 2020). The multimodal hateful meme detection prompts more follow-up research. For example, (Zhou et al., 2021) proposes using a triplet-relation network to improve encoding on texts, images, and captions generated on images. The improved encoding helps final prediction performance. (Pramanick et al., 2021) propose MOMENTA (multimodal framework for detecting harmful memes and their targets) that uses both global and local perspective to detect all kind of hateful memes. In addition, this framework can be easily explainable and can generalize to unseen contexts.

3 Task and Data

The MAMI challenge consists of the two sub-tasks. The task A is a binary classification task on identifying a MEME to be misogynous (labeled as 1) or not (labeled as 0). The task B is a multi-label classification task on identifying a MEME to be non-misogynous or misogyny sub-types that can be overlapped, i.e., *shaming, stereotype, objectification, and violence*. Table 1 reports on counts of 0/1 labels on the five types of labels. Note that on the misogyny label that is the task A's prediction goal, half of MEMEs in the training data are misogynous ($label = 1$). Among four types of misogyny sub-types, we can find that their distributions are not balanced. For example, most frequent sub-type is stereotype with 2810 positive MEMEs while the least frequent label is the violence with only 953 positive MEMEs. For the task A, the evaluation metric is macro-F1. For the task B, the evaluation metric is micro-F1 among five sub-types.

4 Methods

Regarding extracting features from MEME posts' text and image parts, we chose using pre-trained Transformer models to utilize their highly effective feature representations. On texts, we considered two ways, including fine-tuning BERT (Devlin et al., 2019) model and using sentence representations based on BERT, such as Universal Sentence Encoding (USE) embedding (Cer et al., 2018) and SBERT (Reimers and Gurevych, 2019). A major difference between these two ways is that pre-trained BERT model weights are updated in the fine-tuning process. In a contrast, when using USE or SBERT embedding features, these pre-trained models are kept intact.

Regarding the visual encoder processing MEME images to visual representations, we chose a newly emerging Transformer model similar to the BERT model on texts. In recent years, Transformer based visual models have become popular (Han et al., 2020). Among the many visual Transformer models, we selected the ViT model (Dosovitskiy et al., 2020), which is a pure Transformer that is applied directly on an image's $P \times P$ patch sequence. In the implementation, it follows the original Transformer's design as much as possible. ViT utilizes the standard Transformer's encoder part as an image classification feature extractor and adds a MLP head to determine the image labels. The ViT model is pre-trained using a supervised learning task on a massive image data set. The size of the supervised training data set impacts ViT performance significantly. When using Google's in-house JFT 300M image set, ViT can reach a performance superior to other competitive ResNet (He et al., 2016) models. We used the open-sourced pre-trained models on the ImageNet 21K dataset.[1] After converting a MEME image to $P \times P$ patches, ViT converts these patches to visual tokens. After adding a special [CLS] visual token to represent the en-

[1] https://github.com/google-research/vision_transformer

labels	misogynous	stereotype	shaming	objectification	violence
0	5000	8726	7190	7798	9047
1	5000	1274	2810	2202	953

Table 1: Count of label 1 (positive) and 0 (negative) on the misogyny label and other 4 types of sub-types of misogyny in the training set with $n = 10,000$ MEMEs

tire image, the $M = P \times P + 1$ long sequence is fed into a ViT model to output an encoding as $\mathbf{v} = (v_0, v_1, v_2, ...v_M)$, where $M = P \times P$.

CLIP model (Radford et al., 2021) is a seminal work from OpenAI. As shown in Figure 1, on a massive set of image-text pairs, about 350 million, CLIP can pre-train a quite powerful vision-language (VL) joint model by using a simple cross-modal contrastive learning. The trained model shows many impressive applications, like superior performance on many zero-shot image classification tasks. Note that the advantage of using the CLIP model is that the extracted visual and language features have been mapped into a unified embedding space. This will facilitate the next step that combines the two types of features (V for visual features and L for language features) and then uses a simple LR model for predicting misogyny.

After obtaining visual (V) and language (L) features, one solution could be using sophisticated multimodal fusion methods as shown in (Chou et al., 2020) to consider inter-actions between the two types of features. However, after our pilot experiments, we did not find noticeable gains by using such advanced fusion methods compared to the simple *early fusion*, i.e., simply concatenating both V and L features. Therefore, we focused on utilizing the early fusion method in this challenge. Figure 2 depicts our proposed model. The image and text part of a MEME post are sent into the CLIP model to extract both V and L features. Then, the V and L features are combined and fed into a LR model to make a binary prediction on misogyny.

Besides the LR model, we also considered another novel way, DeepInsight (Sharma et al., 2019) that is suggested recently. People were impressed by Convolution Neural Network (CNN) on its universal ability on extracting useful image features. Therefore, DeepInsight was proposed for converting a tabular feature vector into a 2D image and using a CNN model to do classification. On some tasks, such method shows its effectiveness on extracting features from complicate tabular data. In addition, a set of techniques that have shown to be useful for improving image classification performance, such as data augmentation in the training stage, i.e., mix-up (Zhang et al., 2017), or in inference stage, testing time augmentation (TTA), are already developed on the image classification task and can be easily applied.

Model	macro F1
BERT fine-tuning	0.608
SBERT embedding + LR	0.650
USE embedding + LR	0.671
ViT fine-tuning	0.633
visualBERT fine-tuning	0.642
USE,ViT + LR	0.720
CLIP VL features + LR	0.765

Table 2: Macro-F1 on misogyny detection from various models that are based on using pre-trained Transformer models to extract features.

In recent years, a new trend in AI research has emerged and it emphasizes the power brought by data sets (Ng, 2021a). On some AI tasks, the performance increase can be achieved by adding a set of labeled samples and sometimes such new data set could be small. In a contrast, performance increases could be hard to achieve when trying different models. For example, in the challenge (Ng, 2021b), all participants were required to solve the problem by only using data-related methods while keeping using one identical model. We also explored this new approach. In this challenge, we explored the data-centric AI approach by manually annotating more samples in the evaluation set. Although we don't have an access to the coding manual used in the MAMI challenge, we checked the training and trial data that were provided with manual labels. After learning the how-to, we annotated a subset of testing samples and then added these labeled samples into our model's training.

5 Results

Table 2 reports on experiment results on various models based on pre-trained Transformers.

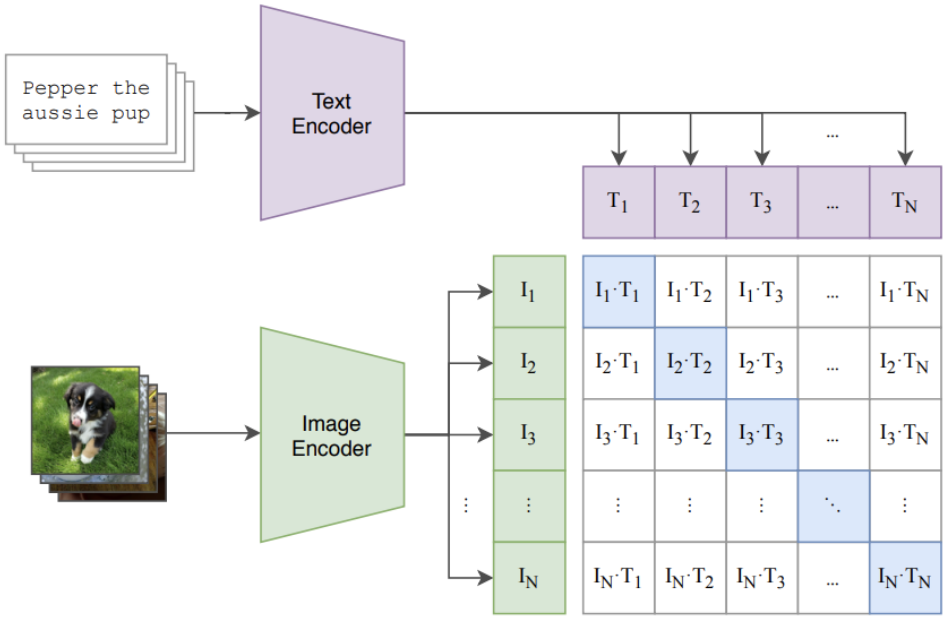

Figure 1: CLIP model is pre-trained on a large number of text-image pairs in a contrastive learning way

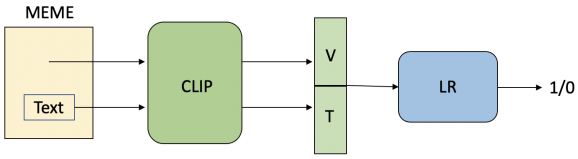

Figure 2: Our model is based on using a logistic regression to detect misogyny by using VL features extracted by the CLIP model

When using BERT fine-tuning, i.e., adding a fully-connected layer on the [CLS] token after BERT model's output layer and fine-tuning two models together by using the cross-entropy (CE) loss, the macro-F1 is 0.608. A different way is obtaining sentence level representations and then feeding these dense features into a LR model implemented in the scikit-learn Python package. We tried both SBERT and USE sentence level representations. This way of using sentence level representations in fact shows improved performance. The macro-F1 can be improved to 0.605 for using the SBERT features and 0.671 for using the USE features.

On image features, we explored ViT model fine-tuning and observed that images play an important role on the misogyny detection. The performance simply using images is 0.633, which is higher than using texts by fine-tuning a BERT model.

Regarding fusing both textual and image features for making a multimodal classification, we tried two methods. The first is fine-tuning a joint Visual-Language (VL) model, visualBERT (Li et al., 2019). However, the performance is not very impressive. Its performance is 0.642, only a slight gain on top of either of uni-modal's performance. The second method is using an *early fusion* by concatenating textual (USE embedding) and image features (ViT embedding) and then fed into a logistic regression (LR) for predicting misogyny. By doing so, the performance can be improved to 0.720.

However, the USE and ViT embeddings are learned from separate models and may not exist in a unified space. To address this issue, We tried the CLIP model for the two attractions, i.e., (1) having coherent textual and visual features in an unified space and (2) image encoder training is on a massive image set with about 350 million images. Consistent to our prediction, after switching to CLIP features, the misogyny prediction's macro-F1 value immediately increased to 0.765.

To explore other possible sophisticated models besides using an LR model, we explored the DeepInsight (Sharma et al., 2019). After converting the dense VL features output from the CLIP model into 2D images, we use a ResNet34 CNN model pre-trained on the ImageNet dataset and converted the misogyny detection into a CNN-based image classification. However, as shown in Table 3, the performance, i.e., 0.751, is worse than using an

Model	macro F1
DeepInsight + CNN	0.751
+ mixup	0.758
+ TTA	0.726

Table 3: Macro-F1 on misogyny detection by using the method converting visual and textual embeddings to 2D images and then using CNN model to do a classification.

Model	macro F1
CLIP VL features + LR	0.765
+ 50 labeled samples	0.767
+ 150 labeled samples	0.772
+ 250 labeled samples.	0.777
using semi-supervised learning	0.778

Table 4: Macro-F1 on misogyny detection by introducing more labeled samples annotated on the test set.

LR model directly. When using the mix-up augmentation, we observed a further performance gain to 0.728. Surprisingly, The TTA method did not show any help. One possible explanation is that we used dense vectors rather than regular tabular data whose feature columns represent some real physical values.

Table 4 reports on the results of utilizing the data-centric AI principle. We can find that by adding increasing number of labeled samples (from the evaluation set), we can keep increasing macro-F1 values. When using labels created by us on 25% of the evaluation set, we can reach a macro-F1 to 0.777. We also used a simple pseudo-label semi-supervised method that is provided by the sci-kit learn Python package and treated all evaluation set ($n = 1,000$) to be unlabeled data. This gave us another small gain to reach our final result of 0.778.

6 Discussions

Multimedia misogyny detection is an important natural language processing application. It uses powerful AI technologies to against misinformation or even harmful information appearing in online communication. For a world emphasizing equal roles between genders, finding misogyny information and removing them is critical for a healthy online communication platform. In this challenge, our methods have been focusing on (a) relying on various pre-trained Transformer models to provide high quality multimodal features and (b) applying the data-centric AI principle to rapidly improve model performances with controllable human efforts. Regarding text encoding, we found that running a simple LR model on top of sentence level representations works consistently better than fine-tuning BERT models. Joint VL model, e.g., visualBERT, does not work quite well in this challenge. However, CLIP, which is trained on a large-sized text-image pair data set in a self-supervised learning approach, can provide high-quality multimodal features and these features can be conveniently used in down-stream classification tasks. On top of highly effective multimodal features, utilizing sophisticated models becomes secondary. In our experiments, simply using an LR model gave us better result than using other complicate models, e.g., DeepInsight. At last, the data-centric AI principle is worth noting. By focusing on our efforts on expanding labeled training data, we can consistently improve our misogyny prediction performance.

References

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

H. Chou, Y.H. Lee, L. Chen, Y. Xia, and W.T. Chen. 2020. CBB-FE, CamemBERT and BiT Feature Extraction for Multimodal Product Classification and Retrieval. In *Proc. SIGIR'20 e-Com workshop*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. ArXiv: 1810.04805.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, and Sylvain Gelly. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 Task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2020. Ami@ evalita2020: Automatic misogyny identification. In *EVALITA*.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018. Overview of the task on automatic misogyny identification at ibereval 2018. *Ibereval@ sepln*, 2150:214–228.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350, Online. Association for Computational Linguistics.

Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, and Yixing Xu. 2020. A Survey on Visual Transformer. *arXiv preprint arXiv:2012.12556*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems*, 33:2611–2624.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Andrew Ng. 2021a. A.i. needs to get past the idea of big data.

Andrew Ng. 2021b. Data-centric ai competition.

Endang Wahyu Pamungkas, Alessandra Teresa Cignarella, Valerio Basile, Viviana Patti, et al. 2018. 14-exlab@ unito for ami at ibereval2018: Exploiting lexical knowledge for detecting misogyny in english and spanish tweets. In *3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages, IberEval 2018*, volume 2150, pages 234–241. CEUR-WS.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Momenta: A multimodal framework for detecting harmful memes and their targets. *arXiv preprint arXiv:2109.05184*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, and Jack Clark. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Alok Sharma, Edwin Vans, Daichi Shigemizu, Keith A Boroevich, and Tatsuhiko Tsunoda. 2019. Deepinsight: A methodology to transform a non-image data to an image for convolution neural network architecture. *Scientific reports*, 9(1):1–7.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Yi Zhou, Zhenhao Chen, and Huiyuan Yang. 2021. Multimodal learning for hateful memes detection. In *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE.

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