taochen at SemEval-2022 Task 5: Multimodal Multitask Learning and Ensemble Learning

Chen Tao[1,2] **and Jung-jae Kim**[1*]
[1]Institute for Infocomm Research, A*STAR, Singapore
[2]Nanyang Techonological University, Singapore
taoc0002@e.ntu.edu.sg, jjkim@i2r.a-star.edu.sg

Abstract

We present a multi-modal deep learning system for the Multimedia Automatic Misogyny Identification (MAMI) challenge, a SemEval task of identifying and classifying misogynistic messages in online memes. We adapt multitask learning for the multimodal subtasks of the MAMI challenge to transfer knowledge among the correlated subtasks. We also leverage on ensemble learning for synergistic integration of models individually trained for the subtasks. We finally discuss about errors of the system to provide useful insights for future work.

1 Introduction

Multimodal machine learning processes data from different modalities (e.g. visual, auditory, lingual) to infer their combined meaning. This topic has seen tremendous development, encompassing visual question answering (VQA) (Stanislaw Antol, 2015), image captioning (Chen et al., 2015), multimodal classification and beyond. SemEval 2022 Task 5 “Multimedia Automatic Misogyny Identification” (MAMI) (Fersini et al., 2022) also requires multimodal machine learning to analyze both visual and textual information from memes (image, caption) in order to identify and classify misogynistic memes. The MAMI challenge has the following two subtasks:

- Subtask A: Classification of a meme as either misogynistic or not
- Subtask B: Categorisation of the type of misogyny if the meme is identified as misogynistic in Subtask A. There are 4 types (or sub-categories) of misogyny: shaming, stereotype, objectification and violence

As introduced by a multimodal research survey (Baltrušaitis et al., 2019), data representation, fusion, and co-learning are the primary challenges in multimodal classification problems. To address those challenges, we explore the following relevant topics: data augmentation, multimodal pre-training, multi-task learning and ensemble learning. Data augmentation helps enrich under-represented features in data. Multimodal pre-trained models may represent multimodal data like those of the MAMI challenge better than unimodal pre-trained models. We also adapt multi-task learning to transfer knowledge in training data among related tasks. As the misogyny sub-categories are correlated to each other, this approach is especially useful in Subtask B. Our final predictions are determined by majority voting among the top performing models.

This paper is structured as follows: In Section 2, we describe the task description and dataset. Section 3 explains in details the models and methods we incorporate into the final system architecture. Section 4 discusses the evaluation results of the models and methods, including multi-task learning and ensemble learning. We conduct error analysis in Section 5 and conclude our findings in Section 6.

2 Task Description and Dataset

SemEval 2022 Task 5 “Multimedia Automatic Misogyny Identification” (MAMI) (Fersini et al., 2022) is a classification task that aims at identifying and classifying misogynistic memes in social media. While most memes are funny and harmless, some deliver misogynistic content and have strong, negative influence due to their high speed of spread. Such memes need to be detected and removed from online sites to avoid gender-related hate. In particular, the MAMI task targets memes, each of which is essentially an image characterized by a pictorial content with an overlaying text a posteriori introduced by human, thus a multi-modal (image+text) analytics task.

The MAMI task has two subtasks. In Subtask A, participants must classify memes into misogynistic and non-misogynistic. The evaluation metric is

[*] Corresponding Author

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 648 - 653
July 14-15, 2022 ©2022 Association for Computational Linguistics

macro-average F1 score. Subtask B is to identify the type of misogyny from the possibly overlapping categories: shaming, stereotype, objectification, and violence if the meme is misogynistic. The evaluation metric is weighted-average F1-Measure, where the F1 scores of the four categories are averaged with weights by support, i.e. the number of true instances for each label.

The training dataset contains 10,000 memes with English captions, assembled from social media platforms. Another 1,000 memes are given as the test dataset. Besides memes in JPEG file format, the task also provides transcriptions of memes captions in a separate text file. Parts of the data are collected and evaluated as described in (Gasparini et al., 2021).

3 Methods

We present a system that utilizes a multi-task learning method to deal with dependencies between the subtasks, fills in the lacking parts of the training dataset by using data augmentation methods, and combines the outputs of multiple base models with ensemble methods.

3.1 Multimodal Base Models

The Hateful Memes Challenge (Kiela et al., 2020b), calling for state-of-the-art models for hate speech detection in multimodal memes, published baseline codes called MMF, which are modular and highly scalable. We thus chose MMF as the codebase of our base models. We extracted 100 features from each meme with the image feature extraction function of MMF and the configurative specifications produced by HateDetectron (Kiela et al., 2020b), the second runner-up of the Hateful Memes Challenge. The feature extractor is backboned by Faster-RCNN (Ren et al., 2015) and ResNet-152 (He et al., 2016).

Below are detailed descriptions of all the base models offered in MMF that we used.

Unimodal (text): We included a unimodal model as a control group. The modality for this model is configurable. After experimentation, we decided to go for text as the single modality for prediction, as it rendered better results than using only images or feature vectors.

ConcatBOW/BERT/MMF: MMF has baseline models for multimodal classification tasks. We obtain text representations from bag-of-word (BOW), BERT, and MMF models and image representations as feature vectors. The "Concat" models concatenate the two representations (text+image) as fusion, and the fusion results are then passed through a Multi-Layer Perceptron (MLP) to output predictions.

LateFusionMMF: The Late fusion model takes the mean between the text and image representations, instead of concatenation.

ViLBERT: Vision-and-Language BERT (ViLBERT) (Lu et al., 2019) is a BERT-based multimodal model pre-trained on the Conceptual Captions (Sharma et al., 2018) dataset, comprising captioned images. It uses two pre-trained strategies: (1) reconstructing image regions or words for masked inputs based on the unmasked portions and (2) prediction of multimodal image-text alignment.

The fundamental difference from the original BERT architecture lies in the attention mechanism, with co-attention used in place of self-attention. By exchanging key-value pairs in multi-headed attention, ViLBERT conducts vision-attended language attention in the visual stream and language-attended vision attention in the linguistic stream. ViLBERT has shown state-of-the-art performance on multiple vision-and-language tasks.

Visual BERT: Like ViLBERT (Li et al., 2019), Visual BERT is another BERT-based multimodal model. It reuses the self-attention system in transformers to implicitly match image regions with texts. The visual embeddings comprise segment embeddings, positional embeddings, and feature vectors of bounding boxes, generated by Faster-RCNN (Ren et al., 2015). They are then fed into the transformer together with text embeddings. With visual and linguistic inputs pre-trained together, the model can discover interesting alignments and implicitly construct effective joint representations.

Visual BERT was pre-trained task-agnostically on the COCO (Chen et al., 2015) dataset for two tasks: masked language modelling and sentence-image prediction. For better adaptation to a particular domain, Visual BERT is usually fine-tuned using masked language modelling on task data before it is applied to downstream tasks.

MMBT: The basic idea of Multimodal Bitransformers (MMBT) (Kiela et al., 2020a) is to apply self-attention to both modalities (i.e. text and image) all at once. This is achieved by utilizing segment embeddings to differentiate between the two modalities, the same method for typical question answering tasks to separate question from paragraph.

A fundamental difference between MMBT and other self-supervised architectures like ViLBERT and VisualBERT is that MMBT only pre-trains individual modalities unimodally. Such design has the plug-and-go advantage if a better vision or language model emerges. It is trivial to replace the pre-trained models for different modalities since they are pre-trained separately. On the other hand, MMBT cannot fully gauge the powerful attention mechanism on multimodal data during pre-training.

3.2 Multi-task Learning

Our baseline models' results (see Section 4) show a significant gap between Subtasks A and B and also uneven performance among the misogynistic sub-categories of Subtask B. This means the models fall short of generalizing the meaning of misogyny. Therefore, we adapted multi-task learning (MTL) to uncover the shared knowledge among all the misogynistic sub-categories.

A MTL learns multiple tasks simultaneously by constructing a generalized representation for the data in different yet related contexts. It is suitable for our dataset since we have five inter-correlated outputs (1 from Subtask A and 4 from Subtask B). To realize multi-task learning, we altered the architecture of the Visual BERT baseline model. While the encoder portion was kept unchanged, we duplicate the classification layer of the decoder into five, each associated with a sigmoid layer for prediction of one of the five outputs. We used binary cross-entropy (BCE) loss combined with Kullback–Leibler divergence loss (KL) as the loss function and summed all the losses generated by individual classification layers. For comparison purpose, we also tried MTL on only the four misogyny types of Subtask B, but the five layers configuration yielded better results.

In fact, we tried to fine-tune Visual BERT on Subtask A first before further fine-tuning it on Subtask B, because we view the misogynistic classification as the main task, and the type categorization task stems from it. Contrary to our expectations, we observed a decrease in model performance. Instead of the misogyny classification inferring the types of misogyny, the results could indicate that the types of misogyny dictate the misogyny classification.

3.3 Data Augmentation

Upon inspection of the validation dataset errors, we found that around 30% of them are blurry or have low resolution, and that about 17% of them express sarcasm against men but are misclassified as misogyny. To address this issue, we created new data by augmenting 10% of the original training set that compare males with females. We augmented them with nine transformations: (1) rotation, (2) Gaussian noise, (3) blurring, (4) horizontal flip, (5) contrast, (6) Affine transformation, (7) distortion, (8) elastic transformation, and (9) change in hue and saturation, with 100 memes per transformation. This boosted the robustness of our model to deal with various data qualities. The data augmentation library we used is *imgaug* (Jung et al., 2020).

3.4 Ensemble Learning

Ensemble learning (Opitz and Maclin, 1999) is an effective method that makes better achievements by combining multiple models' outputs. Taking advantage of the "wisdom of the crowd", an ensemble model can outperform a single contributing model.

Based on our baseline results (see Section 4), all the baseline models except Concat BOW and Unimodal Text have competency on at least one subtask. Thus, the ensemble model pool only excludes the two underperforming models. We experimented with two ensemble strategies, 1) an Multi-layer Perceptron (MLP) and 2) a majority voting layer that averages output probabilities, and compared their results, where both take as input the concatenation of the outputs of the selected baseline models and predict the final outputs per subtask[1].

4 Results

We show the evaluation results of the baseline models, multi-task learning and ensemble methods, and then based on the results, present the overall architecture of the final system we used for our official submission of the MAMI challenge in Section 4.4.

4.1 Baseline comparison

The baseline models serve as the starting point for our experimentation. We trained and evaluated all models included in Section 3.1 individually on all subtasks (misogyny, shaming, stereotype, objectification, violence). Table 1 summarizes the evaluation results.

[1] We refer to the lowercase "subtask" as the misogyny type columns (misogyny, shaming, stereotype, objectification, and violence), in comparison to the capitalized "Subtask" (Subtasks A and B) in the MAMI challenge.

Model	Misogyny	Shaming	Sterotype	Objectification	Violence
Unimodal (text)	80.9%	68.8%	72.2%	74.7%	65.3%
ConcatBOW	74.9%	67.2%	63.5%	73.3%	64.0%
ConcatBERT	**83.7%**	68.0%	71.6%	**77.5%**	**72.0%**
ConcatMMF	**84.4%**	69.6%	**72.7%**	**77.1%**	**74.0%**
Late Fusion	82.8%	**70.9%**	**72.3%**	76.1%	69.1%
ViLBERT	83.4%	69.8%	72.1%	**79.7%**	**75.2%**
VisualBERT	**84.4%**	**72.0%**	**73.8%**	75.5%	71.5%
MMBT	82.3%	**71.4%**	70.9%	73.6%	67.2%

Table 1: Baseline performance evaluated on the validation set. Scores in this table are macro-averaged F1-scores. The best **three** performance for each subtask are highlighted in bold.

Since there was no noticeable change in model performance after we tuned the hyperparameters, we set them as fixed values specified as follows: Enabled early stopping, learning rate of 2e-05, epsilon of 1e-05, batch size of 16, learning rate ratio of 0.6, and warm-up steps of 500.

As seen in Table 1, model performance varies significantly from task to task. Among all the subtasks, Shaming and Violence are the hardest to learn, carrying a non-negligible difference of more than 10% from the Misogyny subtask. This difference in performance will be addressed in the next section.

An interesting note is that early fusion (Snoek et al., 2005) models like Visual BERT are comparable to late fusion (Gunes and Piccardi, 2005) models such as Concat BERT and MMBT. Based on this finding, modal correlations may be picked up at both the decision level as well as the low-dimensional feature level, although the learnt correlations could be different.

Concat BOW performs the worst since it uses simple bags of words as text representation. Its performance is even worse than the unimodal text model. All other multimodal models achieved better results than the unimodal model, showing the involvement of visual information contributes to the overall understanding of the meme contents. We selected the best three models for each subtask for later experiments.

4.2 Multi-task learning evaluation

The alteration of Visual BERT for multi-task learning resulted in an overall advancement of 2.6% for Subtask B and a slight improvement of 1.3% for Subtask A. We also explored different specifications of the classification layers, adjusting the number of hidden layers and activation functions. However, none of them showed significant influence on model's performance.

Method	Subtask A	Subtask B
Top-3 models + MLP	**71.6%**	68.2%
Top-3 models + majority voting	66.1%	**69.5%**
Top-6 models + MLP	70.2%	67.9%
Top-6 models + majority voting	67.4%	69.2%

Table 2: Performance comparison between ensemble learning methods. Scores in this table are of the metrics used for Subtasks A and B, and evaluated on the test set.

4.3 Ensemble evaluation

Table 2 summarizes the evaluation results of two ensemble methods (MLP, majority voting) with the top-*k* models (*k*=3,6)[2]. For each subtask, the top-3 models used are highlighted in bold face in Table 1, and all models except Concat BOW and Unimodal Text are selected as the top-6 models. Refer to Table 1 for the top-3 models selected for each subtask.

The evaluation results of the ensemble methods do not identify a single best method. The majority voting with top-3 models shows the best performance for Subtask B, while MLP is the best for Subtask A. Based on these mixed results, we select different models and ensemble methods for the two Subtasks A and B, which is illustrated in the next section.

4.4 Overall system architecture and evaluation

After considering the evaluation results above, we selected the following three models as parts of the final system architecture:

[2] We refer to the lowercase "subtask" as the misogyny type columns (misogyny, shaming, stereotype, objectification, and violence), in comparison to the capitalized "Subtask" (Subtasks A and B) in the MAMI challenge.

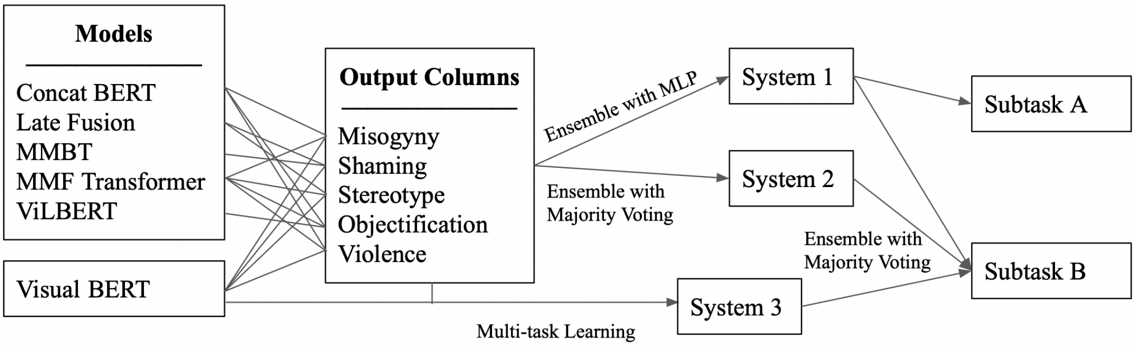

Figure 1: The high-level architecture of the multimodal multitask learning and ensemble learning framework

1. A Visual BERT based multi-task learning model
2. An ensemble learning model that feeds the output probabilities from the top-3 models individually trained on each subtask to a Multi-layer Perceptron (MLP)
3. An ensemble learning model that does majority voting among the top-3 models individually trained on each subtask

In Subtask B, results are generated by applying yet another majority voting layer to these three models. In Subtask A, results are directly taken from the outputs of the ensemble model with MLP (the second model on the list), since we found applying majority voting to Subtask A produced suboptimal results. During the evaluation phase of the competition, we achieved 71.6% in Subtask A and 70.6% in Subtask B, ranked 4th and 6th on the leaderboard, respectively. Figure 1 illustrates how the models are integrated into the final system architecture.

5 Error Analysis

5.1 Analysis on Subtask A

Despite the competitiveness of our system, it still misinterprets the misogyny of some memes in Subtask A. A deeper examination reveals the most common error clusters as discussed below.

First, memes often include references to news, celebrities and cultural practices, e.g. those containing members of "the squad" in the 2019 US House election. This issue may be addressed by detecting entities in memes and collecting their details from the Web to better represent the entities.

Second, the current system falls short of extracting meaning from text positions in memes and the order of sentences (e.g. four-frame mangas). This issue may be addressed by correcting the sentence order and learning their alignment to meme frames.

Third, memes of anti-violence propaganda are often misclassified into the violence category since they include violent-looking images, while the caption is about fighting against violence. In contrast, other misogynistic memes portray violent acts against women in the caption and praise their actions in the image, e.g. with a thumbs-up icon or trophies.

Fourth, memes about comparison between men and women are often confusing. We noticed that these memes either mock males by comparing them to females or describe the real difference between males and females in a funny way. This issue may be addressed by using data augmentation of switching gender-referring words (e.g. man, woman, boy, girl) in the captions of those memes.

5.2 Analysis on Multi-task Learning

In the hold-out set from the training data, our model achieved 83% for Subtask A and 73% for Subtask B. Nevertheless, the same model only yielded 69.7% and 67.5% on the test data. This discrepancy may indicate that training and test data come from different distributions. Conventional multi-task learning is vulnerable to out-of-distribution because it assumes the predicted targets are independent given the input. Unfortunately, the target sub-categories are dependent on each other. This issue might be addressed by adapting, e.g. generative multi-task learning (Makino et al., 2022), which considers the dependency between targets.

6 Conclusion

In this paper, we presented our system for the Multimedia Automatic Misogyny Identification (MAMI) task in SemEval 2022. We augmented the data with visual transformation techniques and extracted features from memes by using multimodal baseline models. We further enhanced the system by adapting multi-task learning and ensemble learning methods. We leave issues such as incorporating external information about entities, frame layout in memes, gender-related text data augmentation and cross-subtask dependency in multi-task learning as future works.

Acknowledgements

We thank Yew Lee Tan for his technical advice. This research is supported by the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funding Scheme (Project #A18A2b0046).

References

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Berta Chulvi Aurora Saibene, 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.

Francesca Gasparini, Giulia Rizzi, Aurora Saibene, and Elisabetta Fersini. 2021. Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content.

H. Gunes and M. Piccardi. 2005. Affect recognition from face and body: early fusion vs. late fusion. In *2005 IEEE International Conference on Systems, Man and Cybernetics*, volume 4, pages 3437–3443 Vol. 4.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Alexander B. Jung, Kentaro Wada, Jon Crall, Satoshi Tanaka, Jake Graving, Christoph Reinders, Sarthak Yadav, Joy Banerjee, Gábor Vecsei, Adam Kraft, Zheng Rui, Jirka Borovec, Christian Vallentin, Semen Zhydenko, Kilian Pfeiffer, Ben Cook, Ismael Fernández, François-Michel De Rainville, Chi-Hung Weng, Abner Ayala-Acevedo, Raphael Meudec, Matias Laporte, et al. 2020. imgaug. `https://github.com/aleju/imgaug`. Online; accessed 01-Feb-2020.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. 2020a. Supervised multimodal bitransformers for classifying images and text.

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

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.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*.

Taro Makino, Krzysztof Geras, and Kyunghyun Cho. 2022. Generative multitask learning mitigates target-causing confounding.

D. Opitz and R. Maclin. 1999. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, 11:169–198.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*.

Cees Snoek, Marcel Worring, and Arnold Smeulders. 2005. Early versus late fusion in semantic video analysis. pages 399–402.

Jiasen Lu Margaret Mitchell Dhruv Batra C Lawrence Zitnick Devi Parikh Stanislaw Antol, Aishwarya Agrawal. 2015. VQA: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.