SRCB at SemEval-2022 Task 5: Pretraining Based Image to Text Late Sequential Fusion System for Multimodal Misogynous Meme Identification

Yujin Wang[*], **Jing Zhang**[✉], **Bohua Peng**[*], **Xudong Zhang**[*]
Xiaoyan Qu, Yimeng Zhuang, Song Liu
Samsung Research China-Beijing (SRC-B)
{yujin1.wang, jing97.zhang, bohua.peng}@samsung.com
{xudong.zhang, xiaoyan11.qu}@samsung.com
{ym.zhuang, s0101.liu}@samsung.com

Abstract

Online misogyny meme detection is an image/text multimodal classification task, the complicated relation of image and text challenges the intelligent system's modality fusion learning capability. In this paper, we investigate the single-stream UNITER and dual-stream CLIP multimodal pretrained models on their capability to handle strong and weakly correlated image/text pairs. The XGBoost classifier with image features extracted by the CLIP model has the highest performance and being robust on domain shift. Based on this, we propose the PBR system, an ensemble system of Pretraining models, Boosting method and Rule-based adjustment, text information is fused into the system using our late sequential fusion scheme. Our system ranks 1st place on both sub-task A and sub-task B of the SemEval-2022 Task 5 Multimedia Automatic Misogyny Identification, with 0.834/0.731 macro F1 scores for sub-task A/B correspondingly.

1 Introduction

Much of the real world's information comes in multimodality, a combination of images, texts, audios and so on. Multimodal understanding aims to utilize different modal of information to improve the overall system recognition intelligence or robustness (Gadzicki et al., 2020), which plays a key foundation role in cognitive AI and embodied AI.

With transfer learning by large deep models and colossal corpus achieving remarkable success in vision and language domain, there is a rising interest in combining both sides' advances to push the multimodality understanding further (Lu et al., 2019; Tan and Bansal, 2019; Chen et al., 2019; Li et al., 2020; Yu et al., 2020; Huo et al., 2021; Kim et al., 2021; Radford et al., 2021). We will limit the discussion scope of multimodal to vision and language in this paper. There are two kinds of representative architecture of multimodal learning models, single-stream models and dual-stream models. Single-stream model fuses the image and text data at an early stage, and then feed into the model. Dual-stream models design separated structure as image encoder and text encoder, and a further module is stacked on top of the unimodel encoders for cross-modal learning objectives (Tan and Bansal, 2019; Yu et al., 2020; Radford et al., 2021; Huo et al., 2021). Usually per-unimodal objectives and multimodal objectives are designed to ensure that the model learns unimodal and crossmodal knowledge, like masked image prediction, masked token prediction, and text-image pairing (Chen et al., 2019; Kim et al., 2021). Two kinds of data distributions are explored for the large-scale pretraining, strongly paired data (Chen et al., 2019; Radford et al., 2021; Li et al., 2020; Kim et al., 2021) and weakly paired data (Huo et al., 2021). The different distributions would directly affect the correlations learned by the model, yet each pretraining corpus only falls in one pattern.

The SemEval-2022 Task 5 (Fersini et al., 2022) Multimedia Automatic Misogyny Identification (MAMI) is a multimodal classification task in English. It targets the identification of misogynous memes (characterized by a pictorial content with an overlaying text a posteriori introduced by human), using the image and text from the meme as input data. It has two sub-tasks: sub-task A: 2-fold classification, to identify whether a meme is misogynous or not; sub-task B: 4-fold fine-grained classification, to further recognize the misogynous meme among potential overlapping categories of stereotype, shaming, objectification and violence.

The relationship of the MAMI paired meme image/text data can vary from highly correlated to weakly correlated or not correlated at all. The semantic logical relationship between the meme's image and text can be: 1) align with each other, containing the same semantic, 2) independent but

[*]Contribution during Intership in Samsung Research China-Beijing.

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 585 - 596

connected to form a complete semantic, 3) irrelevant with each other, only one modality decides the meme's semantic (refer to the appendix for illustrations). In summary this task demands both understanding the image and text, as well as setting up the correct semantic logic between the image/text modalities.

In this paper, we want to investigate either the multimodal pretrained models of different design architecture are capable of handling such complex relationship between vision and language, and whether the pretrain-and-fine-tuning paradigm is advantageous over the feature-extraction-and-machine-learning-classification paradigm. We choose two strong baseline pretrained models, UNITER (Chen et al., 2019) as the single-stream model, and CLIP (Radford et al., 2021) as the dual-stream model, and fine-tuning with cross-entropy softmax is compared against the widely adopted XGBoost (Chen and Guestrin, 2016) classifier. Domain shift is discovered between the train (dev) and test dataset, and the above two paradigm is explored for both in-domain situation and domain shift situation. An adversarial discrimination loss (Tzeng et al., 2015) is added to the fine-tuning deep neural network for domain shift, while the XGBoost classifier is tuned with its hyper-parameters.

Our results show that 1) for both pretrained models, multimodal fine-tuning performs better than unimodal. The CLIP dual-stream model performs sightly better than the UNITER single-stream model on in-domain data, given the much greater pretraining corpus CLIP has than UNITER. On data with domain shift, the CLIP fine-tuning is much more stable than the UNITER model, but both models suffer from great performance degradation. 2) the performance of feature-extraction-and-machine-learning-classification by XGBoost classifier is no weaker than that of fine-tuning on top of pretrained models, the XGBoost classifier utilizing only image features from CLIP hits best performance on the test dataset among all modality combinations, plus that the XGBoost is cheaper to train. 3) for domain shift, the XGBoost classifier is more robust, the domain adversarial loss for fine-tuning brings a small rise, but still falling behind the XGBoost classifier.

Our final winning system is a combination of machine learning and deep neural network, by Pretraining models, Boosting method and Rule-based adjustment, which we name PBR, based on the XGBoost classifier of CLIP image features, and a late sequential fusion of multimodal/text information into the classifier's prediction, followed by rule-based adjustment. The details will be stated in Section 2. Our system gets 0.834 macro F1 score on sub-task A and 0.731 macro F1 score on sub-task B, ranking 1st place in both the tasks in the leaderboard.

2 System Overview

2.1 Overall Architecture

Our 3-stage ensemble system showed in Figure 1 works as following:

- stage 1, the image feature extracted by the CLIP model is learned by the XGBoost classifier, to form a image only prediction. The image/text paired data is used to fine-tune the UNITER model on the MAMI task. And external text datasets together with the MAMI text data is fed into the BERT model to train a text only model on the MAMI task.
- stage 2, the UNITER fine-tuning predictions and the BERT fine-tuning predictions are used to adjust the medium confidence zone of the XGBoost prediction, by our late sequential fusion scheme.
- stage 3, the sub-task A and sub-task B predictions are mutually adjusted, taking advantage of the two sub-tasks' logical inference relationship with each other.

2.2 Deep Pretrained Model for Image and Text Representation

The multi-head attention of transformer architecture modelling the interaction between any two tokens within a sequence by constant $\mathcal{O}(\infty)$ distance, has proved to be powerful in learning deep bidirectional interactions in language and vision (Lu et al., 2019; Dosovitskiy et al., 2020). We choose two transformer based pretrained model to get the image and text representations.

1) **single-stream model UNITER.** UNITER (Chen et al., 2019) is a large-scale pre-trained model for UNiversal Image-TExt Representation. The image and text input are fused early by concatenation, and fed into the transformer module to learn contextualized representations. The pretraining includes unimodal and multimodal tasks. The

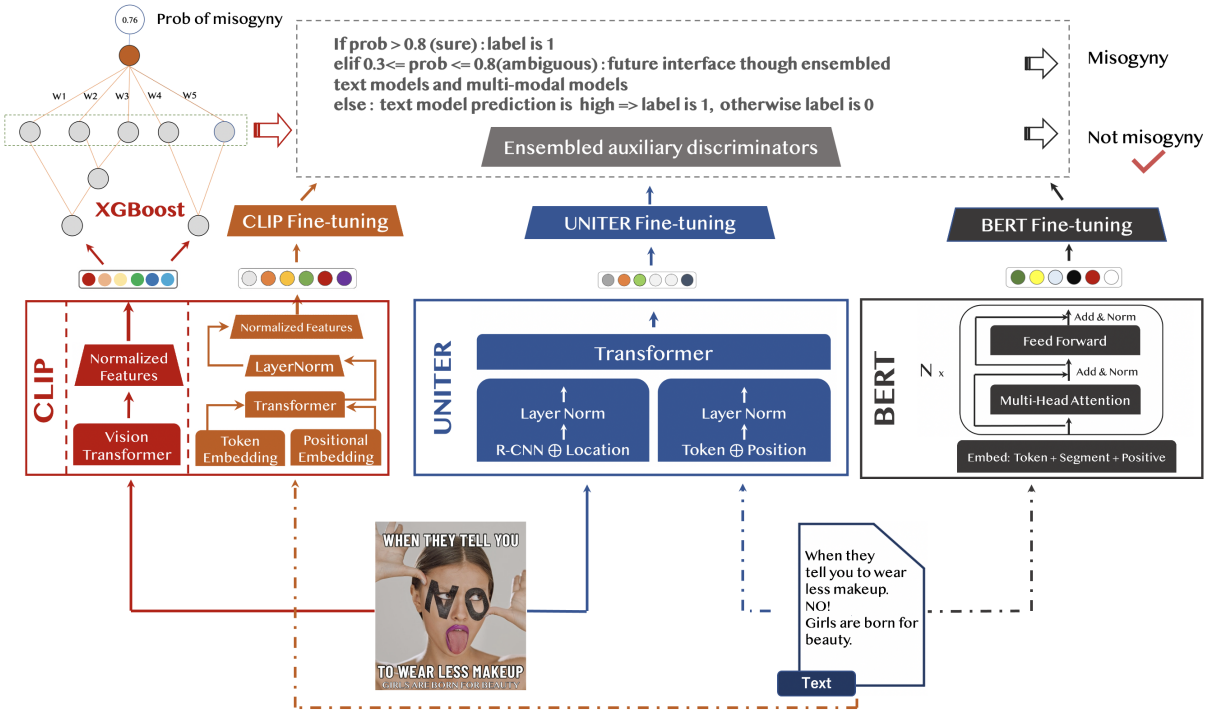


Figure 1: The overall architecture of our ensembled system.

model outputs a fused text and image representation. The pretraining dataset has about 8.4 million image/text pairs.

2) **dual-stream model CLIP.** CLIP (Radford et al., 2021), the Contrastive Language-Image Pre-training model, has separate image transformer encoder and text transformer encoder. The two are joined by a contrastive loss to learn the multi-modal embedding space. The model is pretrained on a dataset of 400 million (image, text) pairs collected from the internet. The simple pretraining task only involves multimodal alignment, predicting which text as a whole is paired with which image, unimodal learning task is not applied. whereas the natural language performs well in enabling zero-shot transfer of the model to downstream tasks when used to reference visual concepts and functioning as prompt text.

2.3 Classification on Downstream Task

2.3.1 Fine-tuning with Pretrained Models

As illustrated in Figure 2, for the UNITER model, the fine-tuning head is a feed forward neural layer (ffn) followed by the cross-entropy softmax classifier. And for the CLIP model, the encoded image and text representations are linearly transformed separately, and then concatenated to be passed forward to a ffn layer and a cross-entropy softmax classifier. The fine-tuning structure of BERT model is the same as the UNITER model. The fine-tuning is used to select the base pretrained model for MAMI, and the fine-tuned UNITER model is utilized as a multimodal voter and the fine-tuned BERT as a text unimodal voter for the model ensemble.

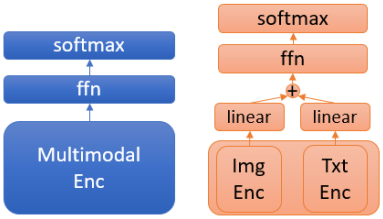


Figure 2: Fine-tuning head structure for UNITER(left) and CLIP(right).

2.3.2 XGBoost Classifier

XGBoost (Chen and Guestrin, 2016) classifier is a tree boosting ensemble model that uses additive functions to predict the output. The boosting ensemble learning algorithm combines multiple weak learners in a sequential method, iteratively improving upon observations. XGBoost borrows from random forests and supports column sampling as well as data sampling. The benefits of the XGBoost classifier is its capability to reduce bias and the low training cost.

In Eq. 1 , the f_k stands for the kth regression

Dataset	Modality	# Samples	# Labels	
		train/dev/test	# misogyny	# non-misogyny
MAMI	img/txt	3227/837/1000	5000	5000
searched-meme	img/txt	3447/-/-	1564	1883
misogynistic-meme	img/txt	800/-/-	400	400
sexist-detection	txt	1142/-/-	627	515
online-misogyny-eacl2021	txt	6567/-/-	699	5868

Table 1: MAMI task and augmented datasets statistics.

tree with K trees in total, $\hat{y}_i$ is the prediction of sample i formed by the sum of K regression trees. In Eq. 2, L is the training objective, l a differentiable convex loss function that measures the difference between the prediction $\hat{y}_i$ and the target y_i, and Ω is the penalty function to avoid over-fitting. The Eq. 3 describes the iterative update of object function L, y_i^t is the prediction of the i-th instance at the t-th interation. At each iteration t, a new tree f_t is added to optimize the objective, the selection of f_t is by a greedy algorithm that most improves the model according to Eq. 2.

$$\hat{y_i} = \sum_{k=1}^{K} f_k(x_i) \tag{1}$$

$$L(\phi) = \sum_i l(y_i, \hat{y_i}) + \sum_k \Omega(f_k) \tag{2}$$

$$L^t(\phi) = \sum_i l(y_i, \hat{y_i}^{t-1} + f_t(x_i)) + \sum_k \Omega(f_k) \tag{3}$$

2.4 Post-adjustification

2.4.1 Late Sequential Fusion

The XGBoost classifier with image features extracted by CLIP stands out among all modality combinations by a large margin, including the multimodal fusion pattern of both CLIP and UNIER, thus it is chosen as our basis. While the information in text is non-negligible, we make use of it in a late sequential fusion fashion, with image going first and text catching up. We treat the XGBoost prediction score ranging [0,1] as the prediction confidence, denoted as $\hat{p}$, and the whole XGBoost predicted confidence on the MAMI test cases are denoted as $\hat{P}$. We rank $\hat{P}$ in decreasing order, and divide it into three intervals, the high, medium and low confidence intervals. The prediction confidence score from the BERT fine-tuning (text modality) and UNITER fine-tuning (multimodality) are denoted as $\hat{p}_b$ and $\hat{p}_u$.

$$\hat{p} = \begin{cases} \hat{p}, & \text{if } \hat{p} \in [\hat{P}_{hi}, \hat{P}_{t_1}] \\ vote(\hat{p}, \hat{p}_u, \hat{p}_b), & \text{if } \hat{p} \in (\hat{P}_{t_1}, \hat{P}_{t_2}] \\ vote(\hat{p}, \hat{p}_b), & \text{if } \hat{p} \in (\hat{P}_{t_2}, \hat{P}_{lo}] \end{cases} \tag{4}$$

The medium confidence interval reflects the model's uncertainty for classification based on image solely. In the low confidence interval, when the image is highly non-misogynous, while the text is highly misogynous, the whole semantic of the meme would be positive. Thus we fuse the text and image modality in a late sequential way by the scheme in Eq. 4. The high confidence interval take the XGBoost classifier prediction as the result directly, and the medium interval combines the text ($\hat{p}_b$) and multimodal ($\hat{p}_u$) information by voting, while the low interval takes advantage of the text information to adjust the image-only prediction. $\hat{P}_{hi}$ and $\hat{P}_{lo}$ refer to the highest and lowest probability, by experience we choose the endpoints of the medium interval t_1 to be 300, and t_2 be 700 in the descending ranked $\hat{P}$ sequence (eg. $\hat{P}_{t_1}$ equals the probability value of the 300th $\hat{P}$).

2.4.2 Mutual Adjustment of the Sub-task A/B

This step is the final adjustment towards our final results. By task definition, when the MAMI subtask A is non-misogynous, all labels in sub-task B should be 0, vice visa. If any of sub-task B is labelled as 1, sub-task A should be misogynous. Therefore, we design the following rules to increase performance.

1) If the prediction of "misogynous" in sub-task A has high confidence for label 0, while some of the four sub-classes in sub-task B are labelled as 1 , we ignore them and set all the labels to 0s due to the high confidence of the misogyny binary classification.

2) If multiple 1-labels appear in sub-task B, it suggests the sample meme probably being misogyny. Meantime if the confidence level of "misogynous" for sub-task A is in the medium interval and the text and visual combination are also very ambiguous, we set the label to be 1 in sub-task A.

2.5 Data Augmentation

Data augmentation is applied to enrich the data distribution and enhance the system's generalization capability (Perez and Wang, 2017) in the downstream task classification phase (applied both in fine-tuning and XGBoost classification).

Our data augmentation strategy includes 3 aspects: 1) collecting memes of the misogyny topic from search engines with a set of misogynous keywords and neutral keywords. 2) collecting public dataset on misogyny and related topics, to help provide more knowledge on the topic. 3) self-augmentation from the task dataset. For image self-augmentation, we used geometric-based augmentations, including flipping horizontally and vertically with cutout (DeVries and Taylor, 2017), randomly resized cropping and 30-degree rotation, as well as color-based transformation, color jittering. For text self-augmentation, back-translation is used. Details of 1) and 2) can be referred in Section 3.1

3 Experimental Setup

3.1 Training Datasets

The MAMI task dataset and augmented datasets are used for training.

MAMI. Dataset for the SemEval-2022 Task 5 (Fersini et al., 2022), the labels of sub-task A is evenly distributed (1:1 for misogyny and non-misogyny samples), and the labels for sub-task B are distributed unevenly, the shaming/stereotype/objectification/violence have 1271/2810/2201/953 labels correspondingly.

searched-meme. Memes crawled from commercial search engines. Searching by keywords in commercial search engines, and an in-house OCR tool is applied to get the paired text for each meme. The final dataset contains 3447 image-text pairs. The searching keywords is listed in the appendix.

misogynistic-meme. An expert-labeled open misogynistic dataset (Gasparini et al., 2021), it contains 800 memes with manually transcribed text, the misogynisticDE field is used as the label for misogyny.

sexist-detection. A text dataset of sexist statements at workplace (Grosz and Conde-Cespedes, 2020), the label for sexism or not is mapped to misogyny or not for the SemEval-2022 Task 5.

online-misogyny-eacl2021. A text dataset of 6567 labels for Reddit posts and comments for online misogyny detection (Guest et al., 2021).

3.2 Training Details

For the fine-tuning of UNITER and CILP, we mainly follow the original paper, detailed hyper-parameters can be referred in the appendix. The hyper-parameters of the XGBoost classifier is listed in Table 5 in the appendix. We treat the sub-task B as four independent binary classification tasks with four independent XGBoost classifiers. Dealing with the label imbalance, we adjust the XGBoost parameter "scale_pos_weight" to achieve good performance. The sub-task A is evaluated using macro-average F1-Measure, the sub-task B is evaluated using weighted-average F1-Measure (Fersini et al., 2022). A point worth noting is that when there is data imbalance of positive (label-1) and negative (label-0) samples, it is more profitable to predict the less labelled ones in the measure of macro F1. So we try to find label-1 in each category of sub-task B as much as possible by tuning the hyper-parameters since label-1 samples are much less than label-0.

4 Results and Discussion

4.1 Multimodal Pretrained Model Selection

We take the BERT text model as the baseline for text modality and the CLIP image model as the baseline for image modality (the image features of CLIP model outperforms state-of-art image pre-trained models in image classification tasks (Radford et al., 2021)), which we will denote as TB and IB below. We randomly split the 10000 training data into train/dev/inner-test data by 8:1:1, and the test data is released by the task organizer. The UNITER image model and text model are tested by putting the unimodal-only data into the model.

As shown in Table 2, the UNITER image model is well below the IB and the UNITER text model is well below the TB on dev dataset. The UNITER image-text multimodal fine-tuning dev results gains large increase compared to its unitmodal implementations, while slightly better than the TB/IB. This suggests that the unimodal pretraining objectives in UNITER is not as well learned as the unimodal

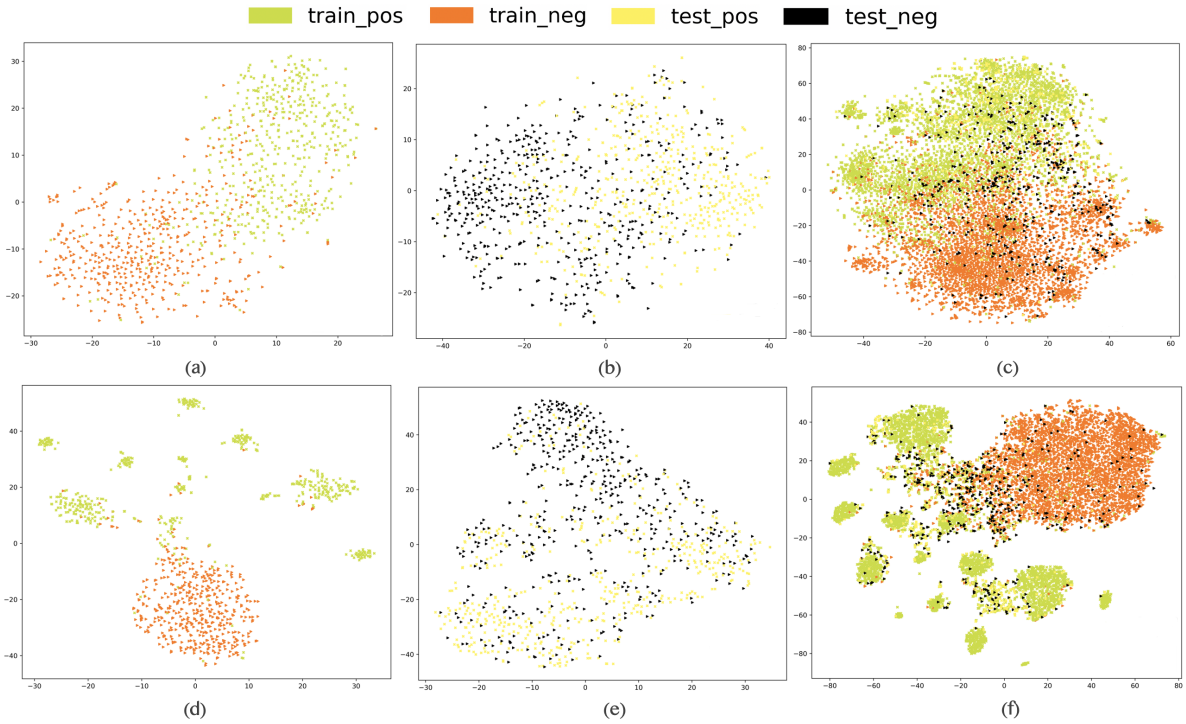


Figure 3: Visualization of t-SNE data distribution under CLIP(a,b,c) and UNITER(d,e,f) models.

Pre-trained model	**Fine-tuning**	
	dev	**test**
BERT txt (TB)	82.6	65.9
UNITER img	70.2	60.3
UNITER txt	76.8	60.7
UNITER img+txt	82.8	67.1
CLIP img (IB)	82.1	68.1
CLIP txt	82.0	66.8
CLIP img+txt	84.3	72.1

Table 2: Baseline performance of single-stream and dual-stream pre-trained models.

benchmarks, and the cross-modal learning objective is better learned through its pretraining given the improvement over the unimodality UNITER models.

The CLIP text model has a comparative performance with the TB, noting that there is no specified text pretraining objective in CLIP. With the well learned image and text unimodal semantics, the CLIP multimodal fine-tuning brings a marinal improvement, leading to 84.3 macro F1 score on dev dataset.

Overall the CLIP multi-modal fine-tuning performs better than the UNITER multimodal fine-tuning, both on dev and test dataset, with comparative unimodal performance at the same time.Thus CLIP is chosen as the pretrained model to provide image/text representations.

4.2 Domain Shift

The big performance gap between the dev and test set in Table 2 suggests domain shift between the train (dev) and test data. Domain shift can be simply expressed as Eq. 5, p_s denotes source data distribution and p_t denotes target data distribution.

$$p_s(x, y) \neq p_t(x, y) \tag{5}$$

$$p(x, y) = p(x|y)p(y) = p(y|x)p(x) \tag{6}$$

According to Bayesian joint probability distribution formula in Eq. 6, the analysis of inconsistent data distributions can be turned to the analysis of marginal probability distributions and conditional probability distributions.

1) $p_s(x) \neq p_t(x), p_s(y|x) = p_t(y|x)$

We consider the training set as the source domain and the test set as the target domain. The distribution of positive and negative samples in the training set is more separable, as shown in Figure 3-a and 3-d. However, the distribution of two classes in test set is mixed, especially the distribution of negative

samples drawn in black triangle shows a significant domain shift,the black triangle, as shown in Figure 3-b and 3-e. Therefore, we regard this as $p_s(x) \neq p_t(x)$ case.

2) $p_s(x) = p_t(x), p_s(y|x) \neq p_t(y|x)$

Another way is to observe the overall distribution in the data set clusters together, as shown in Figure 3-c and 3-f. It is indicated that the distribution of the training set and the test set have relatively low difference. $p_s(x) = p_t(x)$. Generally, the training and test sets are composed of easy samples and hard samples. A hard sample in terms of visualization is a positive sample running into the domain of a negative sample, or a negative sample distributed in the domain of a positive sample. The test set of this competition has a large number of ambiguous samples and difficult samples. This causes the deep learning model to crash, while the more interpretable XGBoost performs better.

To alleviate the problem, we exploit data augmentation and explored further with the XGBoost hyper-parameters

4.2.1 Extra Data for Better Generalization

The external searched-meme dataset and the misogynistic-meme dataset is added for training and the macro F1 score was improved by 2.1 points. This improvement is shown in Figure 4.

4.2.2 XGBoost on Domain Shift

XGBoost has many design parameters to prevent overfitting. These include the number of trees, tree depth, subsampling and colsampling, etc. In addition, there are two penalty terms in XGBoost. $\Omega(f)$ corresponds to Ω in Eq. 2. γ and λ denote the penalty factor, T is the number of leaf nodes. $||w||^2$ is equivalent to the L2 norm in Eq. 7 .

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda||w||^2 \tag{7}$$

As shown in Table 3, the XGBoost classifier shows advantage over fine-tuning on the CLIP model. Different from the fine-tuning paradigm, the image-text CLIP feature performs worst on the XGBoost classifier, with 82.4 macro F1 score. The CLIP text model with XGBoost classifier achieves the best results on dev data, 90.1 macro F1 score, and CLIP image model with XGBoost achieves 85.2 on dev data, but achieves the highest macro F1 score on test data, both of them outperforms the CLIP multimodal fine-tuning. The CLIP image model with XGBoost classifier is chosen as our basis.

Pre-trained model	Fine-tuning dev	Fine-tuning test	XGBoost dev	XGBoost test
CLIP img	82.1	68.1	85.2	**77.6**
CLIP txt	82.0	66.8	90.1	65.4
CLIP img+txt	84.3	**72.1**	82.4	75.1

Table 3: XGBoost Performance on dev and test data compared with pretrained model fine-tuning.

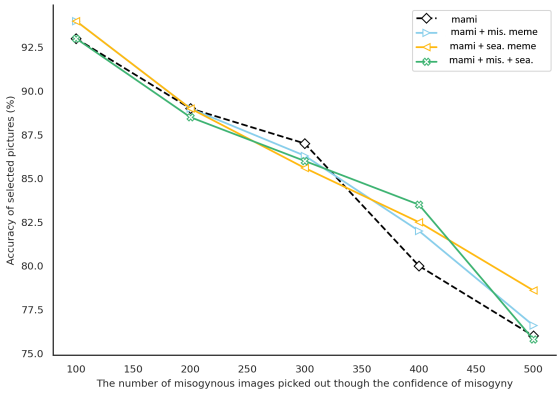


Figure 4: Accuracy of XGBoost classifier trained by different datasets. mis. meme is the misogynistic-meme datase and sea. meme is the searched-meme dataset.

4.2.3 Fine-tuning with Domain Adaptation Compared to the XGBoost Classifier

We apply domain adaption to the CLIP fine-tuning model by an adversarial loss between the source and the target domain following (Tzeng et al., 2015). The core concept is to fuse the distributions of source and target data by a domain classifier together with a domain confusion loss. Besides the standard cross-entropy loss for misogyny classification, the domain classifier (with loss L_{dm}) is on top of CLIP pretrained model to discriminate the source and target data, and the domain confusion loss L_{conf} forces the output of the domain classifier to be a uniform distribution, thus achieving the goal of fusing the source and target domain. By minimising two adversarial losses L_{dm} and L_{conf}, the performance of CLIP model improves by 1.5 on the macro F1 score of testing dataset (72.1->73.6) while it still falls behind the XGBoost.

4.3 Late Sequential Fusion and Mutual Adjustment

The confidence statistics of the XGBoost image (CLIP feature) and BERT text models are shown in Figure 5. It illustrates our intuition of the late sequential fusion. The upper left and lower right

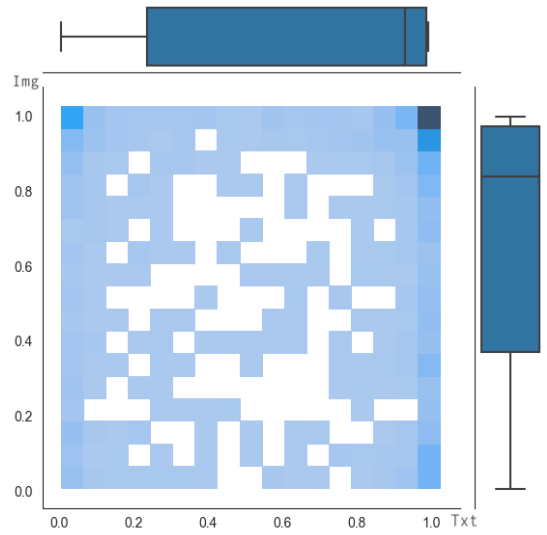


Figure 5: The joint density plot of visual and text modalities on test data.

Task	XG-Boost	Aug. data	Seq. fus.	Mut. adj.
sub-task A	77.6	79.5	81.9	83.4
sub-task B	71.1	-	71.9	73.1

Table 4: Ensemble performance of the system.

corner in the figure shows the disagreement of the image model and the text model. In the medium confidence interval of the XGBoost image prediction, the BERT text prediction can sometimes provide high confidence positive prediction. In addition, the box plot shows the positive skewness distribution in the BERT text model is more obvious than the CLIP-image XGBoost classifier, which means more negative samples are misjudged as positive by the BERT text model. The late sequential fusion boosts the macro F1 score of sub-task A (79.5->81.9) and sub-task B (71.1->71.9). And we get the leading score of 83.4 in sub-task A and 73.1 in sub-task B by the mutual adjustment of the two sub-tasks.

5 Conclusion

In this paper, we investigated the single-steam model UNITER and dual-stream model CLIP's performance on the downstream multimodal classification task, and compared the pretrain-and-fine-tuning paradigm over the feature-extraction-and-machine-learning-classification paradigm.

The experiment results show that the CLIP performs better than UNITER on the MAMI task, and is more robust on domain shift. The UNITER unimodal fine-tuning results are significantly worse than the unimodal pretrain model benchmark, suggesting its weakness in handling the complicated semantic logical relationship in the MAMI task. Wheras the structure of CLIP image feature extraction and XGBoost classificatin achieves the highest baseline performance.

We proposed the late sequential fusion scheme to fuse text information into our system PBR, and exploited extra data and mutual adjustment of the two sub-tasks to further improve the system performance. Our system ranks 1st place in both the sub-tasks in the leaderboard of the SemEval-2022 Task 5 MAMI.

References

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Uniter: Learning universal image-text representations.

Terrance DeVries and Graham W. Taylor. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 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.

Konrad Gadzicki, Razieh Khamsehashari, and Christoph Zetzsche. 2020. Early vs late fusion in multimodal convolutional neural networks. In *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, pages 1–6. IEEE.

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. *arXiv preprint arXiv:2106.08409*.

Dylan Grosz and Patricia Conde-Cespedes. 2020. Automatic detection of sexist statements commonly used at the workplace. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 104–115. Springer.

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.

Yuqi Huo, Manli Zhang, Guangzhen Liu, Haoyu Lu, Yizhao Gao, Guoxing Yang, Jingyuan Wen, Heng Zhang, Baogui Xu, Weihao Zheng, et al. 2021. Wenlan: Bridging vision and language by large-scale multi-modal pre-training. *arXiv preprint arXiv:2103.06561*.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Luis Perez and Jason Wang. 2017. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*.

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

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. 2015. Simultaneous deep transfer across domains and tasks. *arXiv preprint arXiv:1510.02192*.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *arXiv preprint arXiv:2006.16934*.

A Task Data Analysis

The 3 kind of semantic logical relationships between the meme image and meme text in the Introduction Section is illustrated in Figure 9. The subfigures a-c are the cases only text decide the meme's semantic, and d-f are cases that only image decide the semantic, lastly e-g are cases the image and text together form the complete semantic of the meme. Figure 8 shows the high frequency uni- and bigram text distribution. We analyse the top 30 frequent unigram and bigram features of the text input over training and testing distribution (with stop words filtering and ubiquitous words filtering such as "come, "makeameme", "org" which indicate sources of memes). These plots show significant bias, in terms of content and frequency, between train and test distributions.

B Hyper-parameter settings

XGBoost classifier hyper-parameters is shown in Table 5, the BERT fine-tuning hyper-parameters is in Table 6, and the UNITER fine-tuning hyper-parameters shown in Table 7.

Hyper-parameters	Value
objective	binary:logistic
n_estimator	800
learning_rate	0.03
subsample	0.90
max_depth	7
lambda	10
colsample_bytree	0.85
reg_alpha	10
reg_lamba	10
scale_pos_weight	15

Table 5: Hyper-parameters of XGBoost classifier

C keywords for crawling meme data from search engines

The keywords for misogynous memes are {'meme misogyny','meme anti-feminist', 'meme chauvinism woman', 'meme shaming/objectification/stereotype/violence/insult women/woman/girl/female/feminine', 'meme sexist' , 'woman/women/female/feminine hater' }. The keywords for non-misogynous memes are randomly selected neural words like {'meme happy girl', 'meme plants', 'meme school', 'meme actress'} etc.

Hyper-parameter	Value
learning rate	1e-5
learning rate decay	linear
warmup fraction	0.1
Adam ϵ	1e-6
Adam $beta_1$	0.9
Adam $beta_2$	0.98
gradient clip norm	1.0
Weight Decay	0.01
Dropout	0.1
Batch Size	32
Train Epochs	10

Table 6: Hyper-parameters for BERT fine-tuning

Hyper-parameter	Value
learning rate	1e-5
learning rate decay	linear
warmup fraction	0.1
Adam ϵ	1e-6
Adam $beta_1$	0.9
Adam $beta_2$	0.98
gradient clip norm	2.0
Weight Decay	0.01
Dropout	0.1
Batch Size (Token Batch)	5120
Train Epochs	10 for fine-tuning
max txt len	60

Table 7: Hyper-parameters for UNITER fine-tuning

D Experimental results for the late sequential fusion

Figure 6 shows the intial CLIP-image XGBoost classifier's tendency to misclassify the negative sample as positive samples. Figure 7 shows the intermediate ensemble performance.

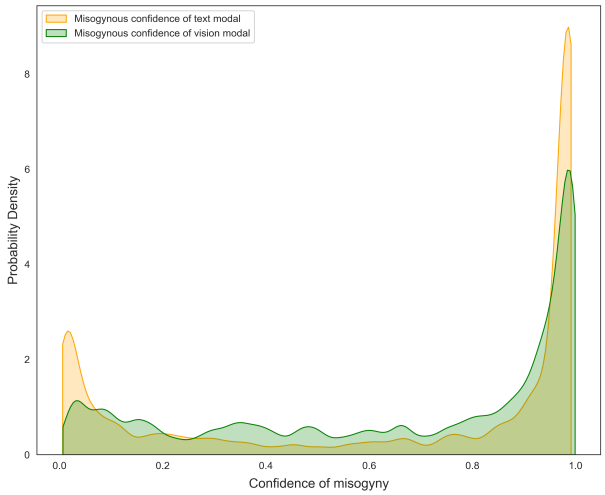


Figure 6: Confidence probability density distribution of textual and visual modalities in "CLIP + XGBoost". Aligned with the TSNE visualization, many negative examples are incorrectly identified as positive examples.

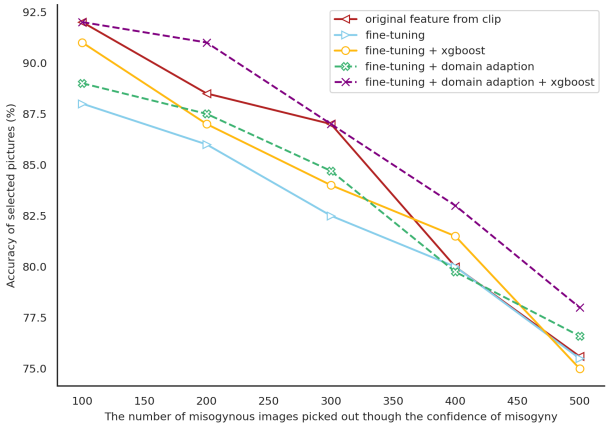
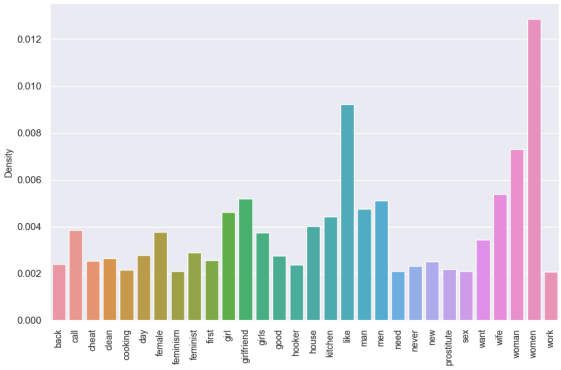
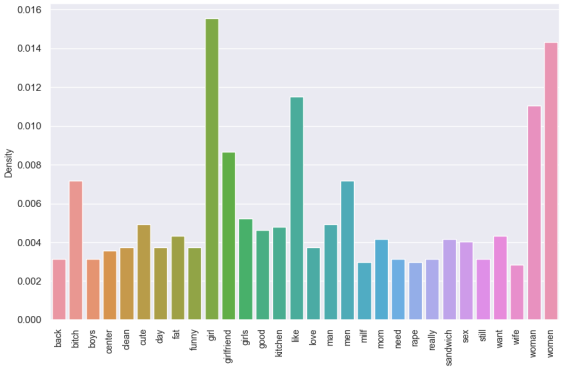


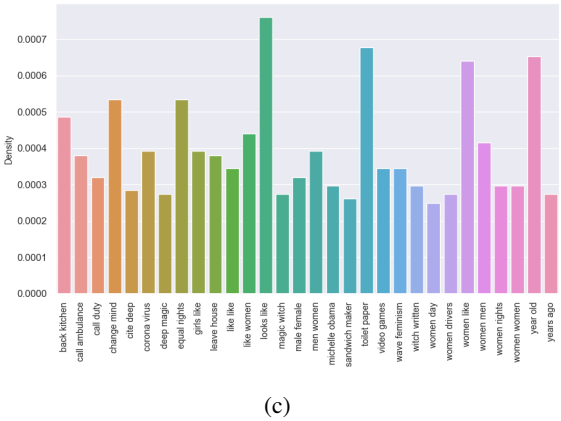
Figure 7: Some extra experimental for the combination of fine-tuning and XGBoost. Normal fine-tuning makes the model learn more towards the training data and performs relatively poorly in the test set. Fine-tuning with domain adaptation can improve the generalization ability of the model. Also according to the dashed line, it can be seen that XGBoot still has a large improvement in the results after fine-tuning.



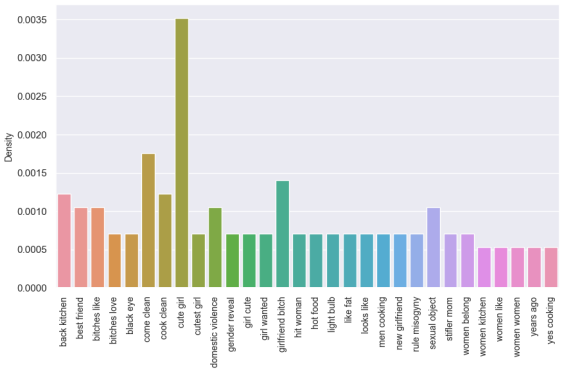
(a)



(b)



(c)



(d)

Figure 8: Text analysis of train dataset and test dataset. (a) and (b) corresponds to unigram features. (c) and (d) corresponds to bigram features of train and test dataset.

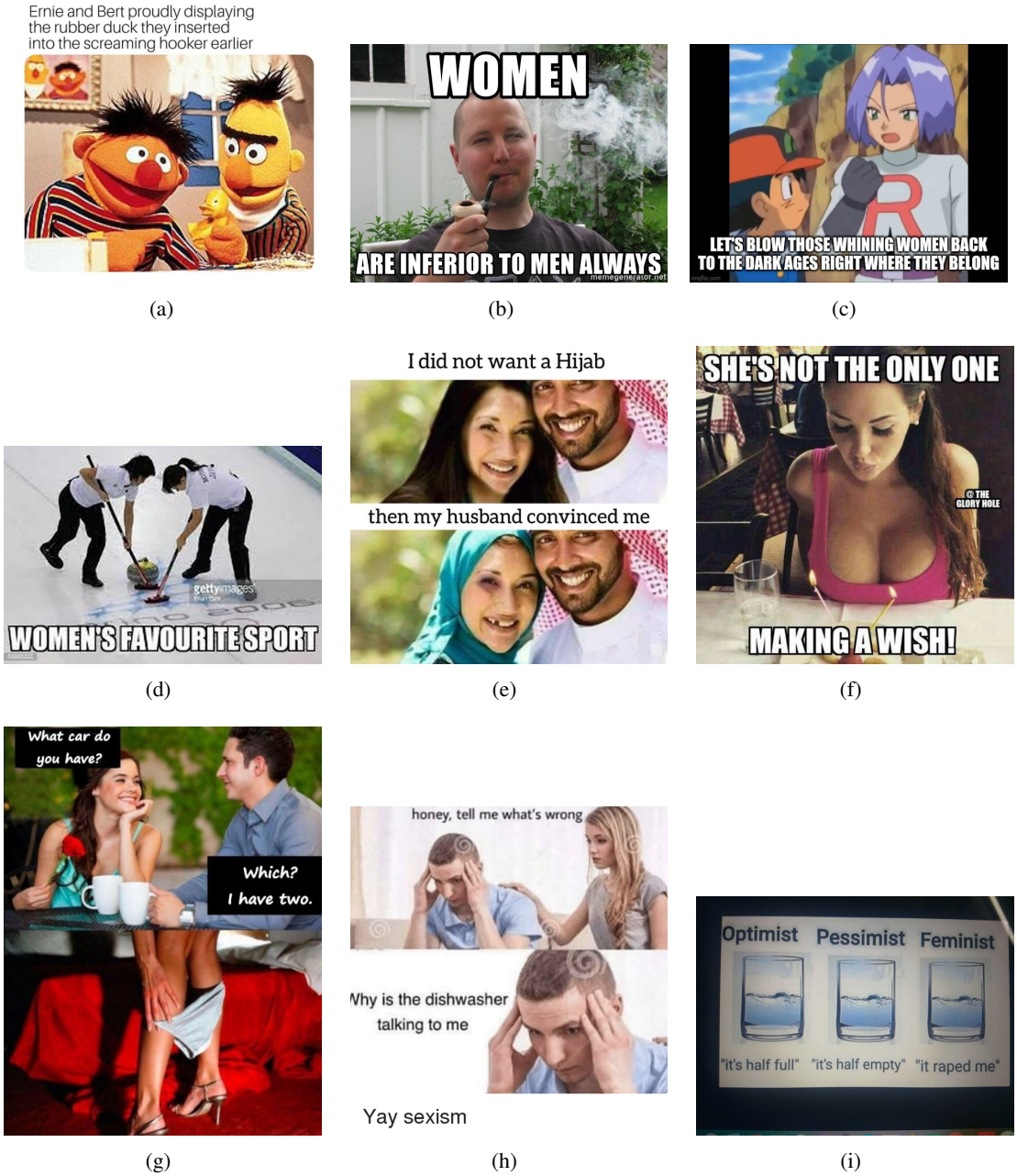


Figure 9: Different image and text semantic relations in MAMI. a-c only text decide the meme semantic, d-f only image decide the meme semantic, g-i text and image together decide the semantic