DD-TIG at SemEval-2022 Task 5: Investigating the Relationships Between Multimodal and Unimodal Information in Misogynous Memes Detection and Classification

Ziming Zhou[2], Han Zhao[1], Jingjing Dong[2], Ning Ding[1], Xiaolong Liu[1], Kangli Zhang[1]
[1]DD-TIG
[2]Peking University
{zhaohan,yaeldingning,xlongliu,zhangkangli}@didiglobal.com
{zhouziming,djj}@stu.pku.edu.cn

Abstract

This paper describes our submission for task 5 Multimedia Automatic Misogyny Identification (MAMI) at SemEval-2022. The task is designed to detect and classify misogynous memes. To utilize both textual and visual information presented in a meme, we investigate several of the most recent visual-language transformer-based multimodal models and choose ERNIE-ViL-Large as our base model. For subtask A, with observations of models' overfitting on unimodal patterns, strategies are proposed to mitigate problems of biased words and template memes. For subtask B, we transform this multi-label problem into a multi-class one and experiment with oversampling and complementary techniques. Our approach places 2nd for subtask A and 5th for subtask B in this competition.

1 Introduction

Online misogynous speech has been a worldwide phenomenon spread widely across social media platforms where women are increasingly subjected to offensive content. It has been shown that women are twice as likely as men to encounter online sexual harassment and gender-based violence (Duggan, 2017).

The problem with misogyny detection is that it requires context and external knowledge to understand online speech, which sometimes can be very short and contain subtle meaning (Kiela et al., 2020). Since memes are getting popular as communication tools on social media platforms, misogynous memes have the potential to affect everyone in our society. Automatic multimodal internet memes identification becomes a new challenging type of misogyny detection task that can only be solved by joint reasoning and understanding of visual and textual information (Zhu, 2020).

The proposed task Multimedia Automatic Misogyny Identification (MAMI) (Fersini et al., 2022) at SemEval-2022 requires participants to identify misogynous memes (subtask A) and classify them as certain overlapping categories: stereotype, shaming, objectification, and violence (subtask B).

This paper describes the system developed by the DD-TIG team for SemEval-2022 Task 5 MAMI. This work contributes to the following: for subtask A, solutions to biased words and template memes are proposed to mitigate the effects of overfitting in unimodal information. We also utilize ensemble learning and external knowledge source like Perspective API to boost the performance of our system. For subtask B, we transform the multi-label classification problem into a multi-class classification problem and reach a better result with oversampling and complementary strategy.

2 Background

2.1 Misogynous memes dataset

MAMI task provides participants with a misogynous memes dataset that contains meme images, the transcriptions of texts on memes, and label annotations. For the training set and test set, misogynous and non-misogynous labels are balanced while misogynous category labels are imbalanced (see Table 1).

Label	Trial	Training	Test
Misogynous	44	5000	500
Non-misogynous	57	5000	500
Shaming	0	1274	146
Stereotype	34	2810	350
Objectification	2	2202	348
Violence	9	953	153

Table 1: Summary of the misogynous memes dataset

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 563 - 570
July 14-15, 2022 ©2022 Association for Computational Linguistics

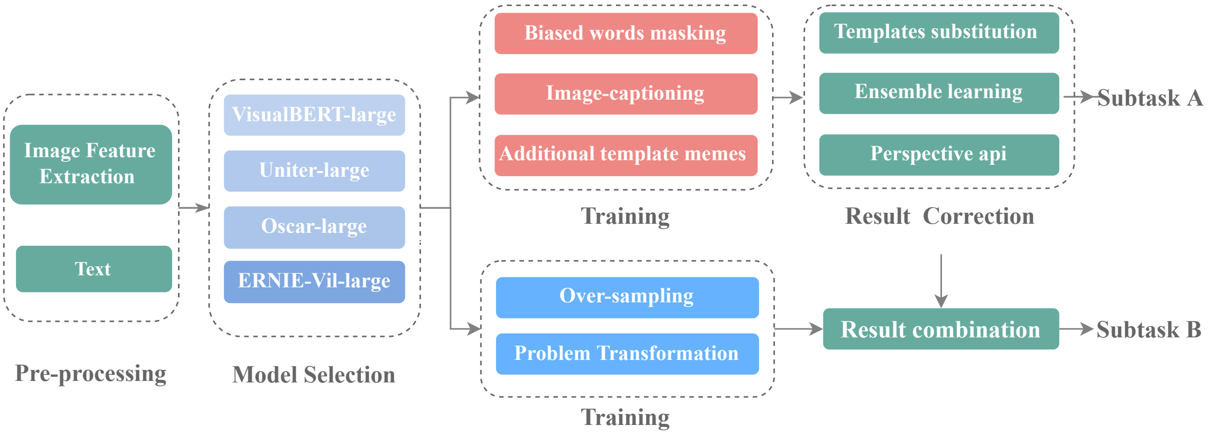

Figure 1: The overall architecture of our proposed system

2.2 Vision and language task

Multimodal misogynous memes identification is a vision and language task. Current state-of-the-art Vision-Language machine learning models are based on the transformer architecture (Vaswani et al., 2017). Among these models, there are two prevalent approaches: single-stream and dual-stream. In single-stream models, such as VisualBERT (Li et al., 2019), UNITER (Chen et al., 2020), OSCAR (Li et al., 2020), image and text features are concatenated and inputted to a standard BERT architecture, which comes under the category of early fusion. In dual-stream models, such as LXMERT (Tan and Bansal, 2019), ERNIE-ViL (Yu et al., 2020a), DeVLBERT (Zhang et al., 2020), VilBERT (Lu et al., 2019), the image and text features are first sent to two independent transformer layers and then into cross-modal transformer layers. Features are combined towards the end of the model as the category of late fusion.

2.3 Vilio: Hateful memes detection framework

The Hateful Memes Challenge (Kiela et al., 2020) is proposed by Facebook AI to leverage machine learning models to solve hateful memes detection problem. Vilio[1] (Muennighoff, 2020) is a code base of 12 different vision+language models and applied to the Hateful Memes Dataset. In our work, we conducted our baseline research on the code of Vilio.

3 System overview

3.1 Preparation

We use the detectron2[2] framework to extract Image features from memes. Detectron2 is provided by Facebook AI with state-of-the-art detection and segmentation algorithms. Specifically, 50 boxes of 2048 dimensions region-based image features are extracted for every meme by Mask-RCNN model. Together with the meme text, which has been extracted using optical character recognition (OCR) and provided in the dataset, features are then fed into the models.

3.2 Vision and language models

We first choose four different base models of VL transformer architectures, namely: VisualBERT, UNITER, OSCAR, and ERNIE-Vil.

We carry out continual pretraining on our dataset with the idea of domain adaptation to reduce the distribution gap between the pretraining dataset and our misogynous memes dataset. MLM pretraining task is taken on pretraining VisualBERT-large, UNITER-large, and OSCAR-large model. However, this does not produce significant performance improvements on our task during the finetuning stage.

Through comparison of results, we found that ERNIE-Vil-large achieves the best performance. In the following steps, we only use the results of ERNIE-Vil-large models.

[1] https://github.com/Muennighoff/vilio

[2] https://github.com/facebookresearch/detectron2

3.3 Strategy for subtask A

3.3.1 Biased words masking

Former research has shown that misogyny detection models can be affected by an unintended bias (Nozza et al., 2019). Some sensitive words, called identity terms, are associated with unreasonably high misogynous scores since they are frequently used in misogynous texts. For example, we observe that the term *kitchen* is frequently used as a stereotypical word against women in our data. Thus, our models tend to associate some non-misogynous texts containing this word with an unreasonably high misogynous score. This situation is known as **unintended bias**, in which models learn usual associations between words (commonly called **identity terms**) which causes them to classify content as misogynous just because it contains one identity word (Godoy and Tommasel, 2021).

Through error analysis on the results of models in the practice stage, we manually collect a list of biased words, including synonyms of woman, dirty words, and controversial words related to feminism. The obtained list of words has been then extended by including their plural form. Refer to Appendix A for the words list.

We propose a novel strategy of biased words masking to mitigate the effects of unintended bias, which also can be regarded as a means of data augmentation. Specifically, when training models, we first use NLTK (Loper and Bird, 2002) to tokenize texts T_i and lemmatize words $w_j \in T_i$ and get $\hat{T_i} = \{w_1, w_2, w_3, \cdots, w_l\}$. Then, for each word w_j in the input text $\hat{T_i}$, with the biased words set $A = \{w_1, w_2, w_3, \cdots, w_k\}$, if $w_j \in A$, it may be masked with a mask token $[mask]$ by a 20 percent probability. We also take strategy with dynamic masking where we generate the masking pattern every time we feed a sequence to the model.

3.3.2 Image captioning

Visual and textual information is semantically aligned in some multimodal tasks, like image-text matching, image-text retrieval, VQA (Yu et al., 2020b). However, for some misogynous memes, image and text are weakly aligned. Thus, there is a semantic gap between visual and textual information. Therefore, we take the strategy of image captioning proposed by previous studies (Das et al., 2020) to enhance model's understanding of visual components. Memes are sent into an image caption model (Xu et al., 2015), which is based on encoder-decoder architecture. This model uses the ResNet-101 as encoder and LSTM as decoder and takes the attention mechanism and beam search when decoding. As a result, this image caption model generates additional descriptions $T_a = \{w_1, w_2, w_3, \cdots, w_i\}$ for visual contents of each meme in the training set; the original text T_o and generated text T_a are concatenated with a separate token $[sep]$.

3.3.3 Templates memes

Through examination of images in the misogyny memes dataset, we notice that many memes are generated by tools of online memes websites. For example, in IMGflip [3], users can choose a meme template from thousands of meme templates and just input their text to caption this template and then get a new meme. In the following part, we refer to memes generated by templates as **template memes**.

In our training set, more than 20 percent of memes are template memes. Some misogynous memes and non-misogynous memes are generated with the same templates and different texts. This may raise a problem that our model associates a high misogynous score or low misogynous score to certain meme templates that actually serve as the medium and contain no misogynous meaning, especially when there is only a few misogynous and non-misogynous sample based on certain templates in the training set. We propose two solutions to mitigate this problem.

Additional template memes: We collect 1,800 memes from memes website. These memes are examples of meme templates and contain no misogynous meaning. Therefore, these memes can be used as the negative sample in our dataset. Without directly adding these memes into the training set, we use our model to make inferences on these memes and only add those false samples (about 90 memes) into the training set.

Templates substitution: we can use an image retrieval model like pre-trained imageNet (Vedaldi and Lenc, 2015) to match memes and templates and find memes that are generated by templates. There are more than 200 memes in the test set are generated with templates. For those memes, original texts and different background pictures are combined to produce K new memes $I_i =$

[3] https://imgflip.com

$\{I_1, I_2, I_3, \cdots, I_k\}$. The probability $\hat{p}$ of a sample will be a combination of model's inferential result on the original meme Ia and new memes I_i with a weighted average w.

$$\hat{p} = \frac{p_a + w \cdot \sum_{j=1}^{k} p_j}{K+1} \tag{1}$$

3.3.4 Ensemble learning

The predicted results of our models can be varying since we take the above-mentioned strategy to train different base models. Thus, we continue to improve the whole system's generalizability and robustness with ensemble learning, where predictions of multiple base models are combined with the method of majority Voting (Velioglu and Rose, 2020). In particular, K (K=20) models are selected for ensemble learning, and predictions are collected from each of the models. The label of data is determined by the majority voted class. We hypothesize that some models show a high recall and low precision and vice versa. So a collection of models may balance out individual weaknesses to achieve better performance than any single model used in the ensemble.

$$\hat{y} = \frac{\sum_{j=1}^{K} y_j}{K} \tag{2}$$

3.3.5 Perspective API

Perspective [4] is a free API that uses machine learning to identify toxic comments, making it easier to host better conversations online. We use Perspective API to get a toxic score for the text of our test data. Labels from the previous models' output and probabilities from Perspective API's results are linearly combined with simple linear regression.

3.4 Strategy for subtask B

3.4.1 Transforming a multi-label problem into multi-class problems

A conventional way to solve a multi-label problem is to transform it into binary classification problems where one binary classifier is independently trained for each label. In machine learning implementation, each unit in the output layer uses the sigmoid activation. This will predict a probability of class membership for the label, a value between 0 and 1. Finally, the model would be fit with the binary cross-entropy loss function. However, there are two problems with this approach. On the one hand, it is troublesome to set an optimal threshold for each label. On the other hand, it does not incorporate information about the relationships between labels. For example, label y_a may only occur by itself; labels y_a and y_b may often occur together; labels y_a and y_c may never occur together.

Since the number of labels in subtask B is 4, which is relatively small, we transform this multi-label problem into multi-class problems. Every possible combination of output labels $([0,0,0,0],[1,0,0,0],\cdots)$ will be taken as a class, and the new space of the label set would be 2^4.

3.4.2 Over-sampling Technique

label	Positive	Negative
Shaming	12.74%	87.26%
Stereotype	28.10%	71.90%
Objectification	22.02%	77.98%
Violence	9.53%	90.47%

Table 2: Distribution of misogynous categories labels in training set

In subtask B, as shown in table 2 , the number of positive samples and negative samples in all misogynous categories is widely imbalanced. Hence, up-sampling of data is done using over-sampling on the positive sample. Thus our new loss function is defined as follows:

$$J = -\sum_{i=1}^{N} \log p_i \cdot \alpha \tag{3}$$

$$\alpha = \begin{cases} \alpha_{neg} & y_i = [0,0,0,0] \\ \alpha_{pos} & otherwise \end{cases} \tag{4}$$

where N is the size of training set; α_{pos} and α_{neg} are the weights for the misogynous and non-misogynous respectively such that $\alpha_{neg} > \alpha_{pos}$ and $\alpha_{neg} + \alpha_{pos} = 1$.

3.4.3 Combination with subtask A results

We train a binary-classification model for task A and multi-class classification for task B separately. Then, we will use the result of Model A to modify the result of Model B, which means if sample X_i a is predicted as $y_a = 0$ or non-misogynous in Model A, it would not belong to any misogynous category, and its predicted label y_b of Model B would be discarded.

$$\hat{y} = \begin{cases} [0,0,0,0] & y_a = 0 \\ y_b & y_a = 1 \end{cases} \tag{5}$$

[4] https://www.perspectiveapi.com

4 Experimental setup

In our baseline approaches, VisualBERT-Large, OSCAR-Large, and UNITER-Large provided by Villo are trained for 5 epochs with a batch size of 16. ERNIE-ViL-Large models provided by the original author are trained for 5000 steps with a batch size of 8. In our work, not much time was spent on hyperparameter optimization, and since we notice that there is not much difference when training models with varying hyperparameter settings and we focus more on other strategies. The hyperparameters for finetuning ERNIE-ViL-Large are presented in Appendix B.

5 Results & Discussion

5.1 Subtask A

Table 3 presents the results of our baseline approaches on the test set, where models are evaluated using Accuracy and a macro-average F1-score, while the latter one is the official metrics for system evaluation in this competition.

Model	Accuracy	F1-score
Oscar-large	69.6	68.9
Uniter-large	69.2	68.4
VisualBERT-large	69.2	68.0
ERNIE-Vil-large	**71.5**	**70.7**

Table 3: The performance of base models on subtask A

ERNIE-Vil has been the STOA model on the multimodal task leaderboard and also achieves competitive performance on our task without any other modification. It is also worth mentioning that continual pretaining with MLM task is conducted on Oscar, Uniter, and VisualBERT, but no improvement is observed. Therefore, ERNIE-Vil is chosen as our base model for further modification with other strategies. Table 4 shows the results of biased word masking, image captioning, and adding false positive samples of template memes into the training set.

Biased word masking experiments have been conducted several times, and the effectiveness is shown by considerable improvement. It is noted that we do not raise scores by only taking the Image captioning technique, but F1-score has a slight increase when image captioning is combined with biased word masking. We hypothesize that image captioning may add noise to our training data since there is a gap between memes in misogynous datasets and the image captioning model's training data. This intuition is confirmed after examining the caption text generated by the Image captioning model, which fails to detect several objects in several images. After we add false-positive samples of template memes into the training set, the performance of our model is boosted. It shows that our model does associate certain normal memes patterns with misogynous or non-misogynous attributes, which can be regarded as biased images.

The results of ensemble learning, templates substitution, and perspective API are shown in Table 4. Ensemble learning obtains a significant improvement where we use different models produced by several times' training of random biased words masking. Since these models are trained with varying texts, we hypothesize their errors will be different, and therefore ensembling may lead to complementary effects and help improve performance. Templates substitution also shows the effectiveness, and this is explained as we find models tend to associate template memes with a high misogynous score, but a majority of them are negative. Perspective API can correct predicted results when sentences contain other malicious words and phrases, but our model does not meet the word or phrase in the training set.

5.2 Subtask B

Table 5 presents the results of our system on the test set for subtask B, where models are evaluated using a weighted-average F1-score, which is the official metric for system evaluation in this competition.

The performance of our model on subtask B is notably improved after the results are modified with the result in subtask A, which has reached a relatively high accuracy score and can be benifical to reduce the number of false positive samples.

5.3 Error analysis

A confusion matrix for subtask A (see Table 6) and a classification report for subtask B (see Table 7) are presented, which will be combined with some bad cases to have both qualitative and quantitative assessments on our system.

	Misogynous	Non-misogynous
Misogynous	328	46
Non-misogynous	172	454

Table 6: Confused matrix for subtask A

For subtask A, obviously, the problem with

Model	Accuracy	F1-score
ERNIE-Vil-large	71.5	70.7
ERNIE-Vil-large + WM	72.8	72.1
ERNIE-Vil-large + IC	71.0	70.6
ERNIE-Vil-large + WM + IC	72.7	72.5
ERNIE-Vil-large + WM + IC + AD	73.8	73.7
ERNIE-Vil-large + WM + IC + AD + Emsembling	76.7	76.5
ERNIE-Vil-large + WM + IC + AD + Emsembling + TS	78.1	78.0
ERNIE-Vil-large + WM + IC + AD + Emsembling + TS + PA	**79.4**	**79.3**

Table 4: The performance of our systems on subtask A (WM is biased words masking. IC is image caption. AD is addtional data of template memes. TS is template substitution. PA is the Perspective API.)

Model	F1-score
ERNIE-Vil-large	70.8
ERNIE-Vil-large + Oversampling	71.3
ERNIE-Vil-large + Oversampling + PT	71.7
ERNIE-Vil-large + Oversampling + PT + RC	**72.8**

Table 5: The performance of our systems on subtask B (PT is problems transformation into multi-class classification. RC is results combination with subtask A)

our system is that a considerable number of non-misogynous samples, about 17.2 percent of total and 34.4 percent of the negative sample, is misclassified as misogynous, as the error type false positive. The goal of strategies like biased words masking is to reduce the effects of certain patterns in texts or images and prevent overfitting. Some patterns still are regarded as crucial features of misogyny by our models, but actually, they are biased. Enlarging the size of our dataset may be beneficial to deal with this problem.

The figure in Appendix C shows an example labeled as non-misogynous in the dataset but predicted as misogynous by our model. As mentioned in the previous part, the word kitchen frequently appears in the misogynous sample since misogynists always hold the stereotype to associate women with certain gender roles. We try to mitigate the bias by biased word masking, but it still can not be solved. Moreover, a girl in this image may be associated with the text, but the image and text are not aligned in fact.

label	Precision	Recall	F1-score
Shaming	0.36	0.55	0.43
Stereotype	0.62	0.64	0.63
Objectification	0.69	0.70	0.69
Violence	0.64	0.55	0.59

Table 7: Classification report for subtask B

For subtask B, our model shows relatively poor performance on the label shaming. According to the definition of shaming provided by MAMI organizers, a shaming meme aims at insulting and offending women because of some characteristics of the body. There are two possible reasons to explain this problem. First, we notice that there are some female characters in memes generated by mocking templates, and in truth, the texts on the memes are not targeted towards the female characters in the memes. Second, the definition of shaming is vague and overlaps with other categories of misogyny.

Thus, there is the challenge of this competition: the information from the image and text modalities should not always be treated equally. Sometimes text information should be emphasized if this meme is based on some templates. In multimodal understanding and reasoning tasks, unimodal information can be imbalanced.

6 Conclusion

In this paper, we have presented our work on Multimedia Automatic Misogyny Identification (MAMI) at SemEval-2022. Mainstream vision-language models are applied on misogynous memes dataset in the baseline approach. For subtask A, to better utilize multimodel information and unimodal information, we propose solutions to mitigate the effects of biased words and templates memes. Ensemble learning and external knowledge source like perspective API are used to enhance the performance of our system. For subtask B, training with oversampling strategy, we use a multi-class model to solve this multi-label problem and gain improvement from our subtask A model. In short, this task could never be solved easily since it relies heavily on the context, external knowledge, relations between modalities.

References

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Abhishek Das, Japsimar Singh Wahi, and Siyao Li. 2020. Detecting hate speech in multi-modal memes. *arXiv preprint arXiv:2012.14891*.

Maeve Duggan. 2017. Online harassment 2017.

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.

Daniela Godoy and Antonela Tommasel. 2021. Is my model biased? exploring unintended bias in misogyny detection tasks.

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*.

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.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

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.

Niklas Muennighoff. 2020. Vilio: state-of-the-art visio-linguistic models applied to hateful memes. *arXiv preprint arXiv:2012.07788*.

Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. 2019. Unintended bias in misogyny detection. In *Ieee/wic/acm international conference on web intelligence*, pages 149–155.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Andrea Vedaldi and Karel Lenc. 2015. Matconvnet: Convolutional neural networks for matlab. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 689–692.

Riza Velioglu and Jewgeni Rose. 2020. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge. *arXiv preprint arXiv:2012.12975*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

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

Zhou Yu, Yuhao Cui, Jun Yu, Meng Wang, Dacheng Tao, and Qi Tian. 2020b. Deep multimodal neural architecture search. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3743–3752.

Shengyu Zhang, Tan Jiang, Tan Wang, Kun Kuang, Zhou Zhao, Jianke Zhu, Jin Yu, Hongxia Yang, and Fei Wu. 2020. Devlbert: Learning deconfounded visio-linguistic representations. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4373–4382.

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

A Biased words list

bitch, bitches, clean, cooking, dish, equal, female, females, feminist, feminists, fuck, fucking, gender, genders, hooker, hookers, horny, house, housewife, kitchen, mama, mom, moms, prostitute, prostitutes, sex, sexism, sexual, single, wash

B Hyperparameters setting

	ERNIE-ViL-Large
Training steps	5000
Warm steps	500
Learning rate	1e-5
Learning rate decay	0.1
Batch size	8
Fusion method	sum
Attention dropout	0.1
Dropout rate	0.5
Max seqence length	256
Optimizer	AdamW

Table 8: Hyperparameters setting for finetuning ERNIE-ViL-Large

C An example of bad cases

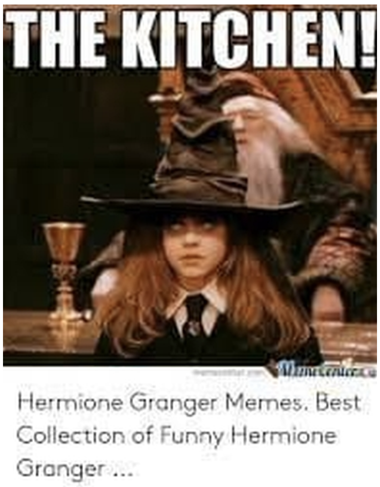
THE KITCHEN!
Hermione Granger Memes. Best Collection of Funny Hermione Granger ...

Figure 2: An example of bad cases