YMAI at SemEval-2022 Task 5: Detecting Misogyny in Memes using VisualBERT and MMBT MultiModal Pre-trained Models

Mohammad Habash **Yahya Daqour** **Malak Abdullah** **Mahmoud Al-Ayyoub**
Computer Science Department
Jordan University of Science and Technology
Irbid, Jordan
{mshabash187,yidaqour18}@cit.just.edu.jo, {mabdullah,maalshbool}@just.edu.jo

Abstract

This paper presents a deep learning system that contends at SemEval-2022 Task 5. The goal is to detect the existence of misogynous memes in sub-task A. At the same time, the advanced multi-label sub-task B categorizes the misogyny of misogynous memes into one of four types: stereotype, shaming, objectification, and violence. The Ensemble technique has been used for three multi-modal deep learning models: two MMBT models and VisualBERT. Our proposed system ranked 17th place out of 83 participant teams with an F1-score of 0.722 in sub-task A, which shows a significant performance improvement over the baseline model's F1-score of 0.65.

1 Introduction

Participating in online social networks (OSN) has become more common nowadays, especially for women, as 78 percent of them use social media several times per day compared to 65 percent of the men. Anyhow, hateful speech against women did not stop until offline methods but also found its way on the web using popular communication tools such as memes. A meme is an online spread of captioned pictures or GIFs meant to be funny or critical to people or society. It has the most humor characteristics on social network platforms. However, in a short period, some minorities started to develop memes aiming at gender discrimination, inequality, and spreading hate against women.

Due to the growth of Artificial Intelligence (AI), Natural Language Processing (NLP), and Computer Vision (CV), the machine started to recognize and understand languages and images more than ever. Therefore, preventing hate speech and misogyny became more achievable by applying these methodologies. This paperwork presents our participation in SemEval-2022 Task 5 (Elisabetta Fersini, 2022) which aims to detect misogyny in memes. In this study, MultiModal BiTransformers (MMBT) (Kiela et al., 2019) and VisualBERT (Li et al., 2019) have been used to build the proposed model by employing the ensemble technique. The proposed model reaches an F1-score of 0.722 in sub-task A depending on the provided 10K memes dataset and without external datasets.

This paperwork is constructed as follows: Section 2 presents the related work, followed by Section 3, which clarifies the task description. Section 4 shows dataset details and preprocessing procedures. Section 5 shows off the architecture of the solution system. Finally, sections 6, 7, and 8 provide the experiments, results, and conclusion.

2 Related Work

Several researchers are attracted to detecting hate, harm, sarcasm, and satire on Social media (Faraj and Abdullah, 2021; Isaksen and Gambäck, 2020; Watanabe et al., 2018). Internet Memes are considered one of the most popular ways to communicate on all topics on social media. Since it is a global issue and a world trend, many researchers have started to compete in preventing harmful memes from pervasion. This paperwork (Shang et al., 2021) aimed to detect offensive analogy memes to prevent this from spreading out. The focus was here on images as it attracts more people and has richer information than the text alone.

One competition from Facebook (Kiela et al., 2021) aimed to detect hate speech in memes as it is hard to be tackled by having humans checking out every meme. From their vision, hate speech was understood as "any communication that disparages a target group of people based on some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics" (Levy et al., 2000). The dataset consisted of

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 780 - 784
July 14-15, 2022 ©2022 Association for Computational Linguistics

12540 memes, split between 7834 for non-hate and 4706 for harmful.

On the other hand, deep learning models have been successfully applied to learning features for single modalities: text (Abdullah and Shaikh, 2018; Abedalla et al., 2019) and images (Abedalla et al., 2021, 2020). Moreover, learning-based models that combine elements from multiple modalities are more robust in visual-linguistic tasks such as detecting hateful memes (Sandulescu, 2020).

3 Task Description

The main objective of this task (Elisabetta Fersini, 2022) is the identification of misogynous memes. Memes can be identified using text, images, or both. Using both images and text surely produces better results. The task is divided into two sub-tasks:

- Sub-task: This task aims to classify the meme as either misogynous or not misogynous.
- Sub-task B: this is a multi-label task, and the goal is to identify the type(s) of misogyny in misogynous memes. Types of misogyny are stereotypes, shaming, objectification, and violence.

Although we did not use any external datasets, it is allowed for participants to do so.

4 Dataset

4.1 Dataset Description

The training dataset provided by (Elisabetta Fersini, 2022) consists of 10,000 memes along with a file containing the extracted text, image file names, and five columns representing labels. The evaluation dataset consists of 1,000 memes, provides the extracted text and does not contain the labels. Competitors are supposed to use the evaluation dataset to produce and submit evaluation results. Memes were gathered by searching on social media platforms, such as Twitter and Reddit, and other meme creation and sharing websites, such as 9GAG, Knowyourmeme, and Imgur.

4.2 Data Preprocessing

Regarding the extracted text of memes, it is considered clean, and there is not much pre-processing required, but a lot of memes contain a web address in the bottom-right corner, which might better be removed. Also, Twitter hashtags and usernames are removed. Furthermore, any unwanted non-English or non-understandable characters are removed. The previously mentioned pre-processing steps provide a slight performance improvement. Regarding the images, no pre-processing is applied.

Statistical Property	Value
Maximum sentence length	325
Minimum. sentence length	2
Average sentence length	21.46
Standard deviation of sentence length	16.97

Table 1: Statistical properties of memes text

4.3 Data Analysis

The dataset contains 5,000 misogynous memes and 5,000 non-misogynous memes. Thus, the dataset is perfectly balanced for sub-task (A). As shown in Table 1, the longest sequence of memes text is 325 words long, and the shortest sequence is only two words long. The average sentence length is 21.46, and the standard deviation of sentence lengths is 16.97. The previous statistics indicate that most sequences are nearly 2-38 words long, and fewer sentences contain more than 38 words. Therefore, the chosen threshold for sequence length (maximum sequence length) is 64.

5 Systems Description

5.1 Text-based Model

The first approach uses pre-trained language models to identify misogynous memes based on text. We ended up choosing RoBERTa (Liu et al., 2019) as it is one of the state-of-the-art language models. Unfortunately, relying on text-only is not good enough to classify memes robustly.

5.2 Image-based Model

The second approach uses pre-trained computer vision models to identify misogynous memes based on images. The chosen model is VGG (Simonyan and Zisserman, 2014), it is one of the most popular Convolutional Neural Network (CNN) models, and the chosen variant is VGG-16 which consists of 16 layers. Unfortunately, similar to text, using images only does not classify memes robustly.

5.3 Multimodal Models

As memes contain both text and images, it is reasonable to use deep learning models that use both text and images to produce more accurate and robust predictions.

The first model used in our solution is Multimodal Bi-Transformers (MMBT) (Kiela et al., 2019). MMBT was developed based on the transformer architecture. It uses the attention modules of the transformer to combine embeddings of different modalities (text and image in our case). The key approach used in MMBT is to take the image as an input, extract its features using a CNN model, concatenate the generated features with the text input tokens used in BERT (Devlin et al., 2018), and feed the model with the text and image tokens. We used ResNet-152 (He et al., 2016) as the CNN model for feature extraction (image encoding) in MMBT.

The second model in our solution is VisualBERT (Li et al., 2019) which is built to fulfill a wide range of vision and language tasks. It consists of a stack of transformer layers similar to BERT architecture to prepare embeddings for image-text pairs. BERT tokenizer is used as a text encoder. For images, a custom pre-trained object detector must be used to extract regions and bounding boxes fed to the model as visual embeddings. We used Detectron2 (Wu et al., 2019) to generate the visual embeddings using MaskRCNN+ResNet-101+FPN model checkpoint.

As shown in Figure 1, The final model is constructed using the voting technique between two MMBT models and the VisualBERT model. Both MMBT models use ResNet-152 for image encoding, but they differ slightly as one uses BERT-base-uncased for the text while the other uses BERT-large-uncased. It is worth mentioning that the two MMBT models were trained using different random seeds.

6 Experiments

Regarding sub-task A, we have experimented with RoBERTa as it is one of the state-of-the-art language models. We have used Huggingface (Wolf et al., 2019) and Pytorch (Paszke et al., 2019) to implement roberta-base. We chose a maximum sequence length of 64. Following this, we have experimented with VGG-16, which is one of the most popular pre-trained CNN models, and implemented it using Keras (Chollet, 2015). We set the model to expect images of size 300x300 and added a dropout layer with a drop rate of 0.3, followed by a dense layer. For RoBERTa and VGG-16, we split the provided dataset into training and validation datasets with a validation split value of 0.1. Both RoBERTa and VGG-16 performed poorly, as memes classification tasks require models that understand images and text.

Model	RoBERTa	VGG-16	MMBT	VisualBERT
Epochs	2 5	50	5 10	5 10
Batch Size	32 64	32 64	16 32 64	32 64
Learning Rate	2e-5 5e-5	0.001	1e-5 2e-5 5e-5	3e-5 5e-5
Weight Decay	0	0	0.0001 0.001	0.001
Gradient Acc. Steps	1	1	1 2	1
Random Seed	17	17	1337 17	42

Table 2: Hyper-parameters of all models for sub-task A.

Subsequently, we have experimented with two vision+language models: MMBT and VisualBERT. Both of them were implemented using Huggingface and Pytorch. MMBT was experimented with using ResNet-152 as image encoder and BERT as text encoder. Two BERT variants were used: BERT-base-uncased and BERT-base-large. Visual embeddings for visualBERT were generated using MaskRCNN+ResNet-101+FPN model checkpoint from Detectron2. For MMBT and VisualBERT, we used k-fold cross-validation with a k value of 5. In other words, the training and validation datasets consist of 8000 and 2000 memes, respectively, and after each epoch, the validation dataset switches to a completely different 2000 memes. This allows the models to train on the entire dataset eventually. In our experiments, MMBT outperformed VisualBERT. AdamW optimizer was used for all models. Table 2 shows different sets of hyper-parameters used for all previously mentioned models.

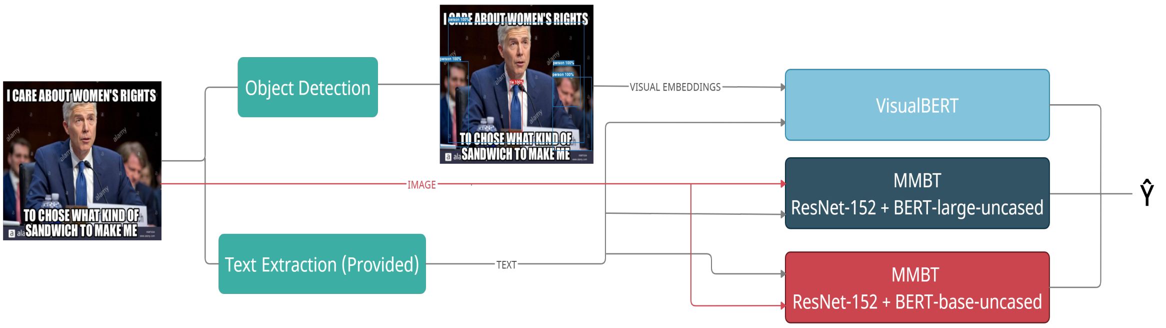
Object Detection
I CARE ABOUT WOMEN'S RIGHTS
TO CHOSE WHAT KIND OF SANDWICH TO MAKE ME
VISUAL EMBEDDINGS
IMAGE
TEXT
Text Extraction (Provided)
VisualBERT
MMBT ResNet-152 + BERT-large-uncased
MMBT ResNet-152 + BERT-base-uncased
Ŷ

Figure 1: The architecture of the final system for sub-task A. ŷ represents the final prediction after applying voting technique between models. The pipeline also illustrates some pre-processing steps before modelling.

7 Results

The final results for sub-task A are produced using the ensemble technique between two MMBT models trained using different random seeds and VisualBERT. RoBERTa and VGG-16 were not included in the final system as they performed poorly compared to MMBT and VisualBERT. As shown in Table 3, both MMBT models slightly outperformed VisualBERT, but the latter still has a positive effect on the final result after the ensemble.

Looking at Figure 2, which shows the confusion matrix for the final results in Sub-task A, We can see that the final system succeeded in predicting 359 non-misogynous memes and 363 misogynous memes. The system also failed to identify 137 misogynous memes and mistakenly identified 141 non-misogynous memes as misogynous.

Model	MMBT (ResNet + Bert-base)	MMBT (ResNet + Bert-large)	VisualBERT	Ensemble
Epochs	5	5	5	*
Batch Size	32	32	32	*
Learning Rate	1e-5	1e-5	5e-5	*
Weight Decay	0	0	0.001	*
Gradient Acc. Steps	1	1	1	*
Random Seed	1337	17	42	*
F1-Score	0.695	0.697	0.679	0.722

Table 3: Final experiments used in sub-task A. The final F1-score of 0.722 is obtained by using ensemble technique between the three models.

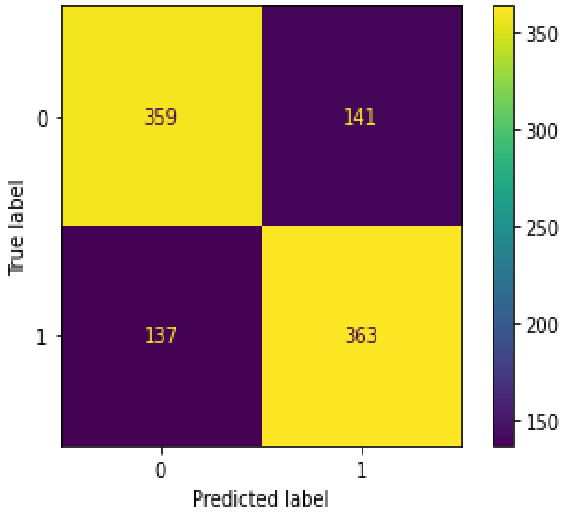

Figure 2: Confusion matrix for the final results in sub-task A

8 Conclusion

This paper presented our deep learning system that contended at SemEval-2022 Task 5. We experimented with different deep learning models, starting with the RoBERTa language model and the VGG-16 CNN model. Subsequently, we experimented with vision and language models, such as MMBT and VisualBERT, which significantly outperformed VGG-16 and RoBERTa. Using ensemble technique between two MMBT models and VisualBERT produced an F1-score of 0.7222 which led to ranking 17th place in sub-task A.

References

Malak Abdullah and Samira Shaikh. 2018. Teamuncc at semeval-2018 task 1: Emotion detection in english and arabic tweets using deep learning. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 350–357.

Ayat Abedalla, Malak Abdullah, Mahmoud Al-Ayyoub, and Elhadj Benkhelifa. 2020. 2st-unet: 2-stage training model using u-net for pneumothorax segmentation in chest x-rays. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE.

Ayat Abedalla, Malak Abdullah, Mahmoud Al-Ayyoub, and Elhadj Benkhelifa. 2021. Chest x-ray pneumothorax segmentation using u-net with efficientnet and resnet architectures. *PeerJ Computer Science*, 7:e607.

Ayat Abedalla, Aisha Al-Sadi, and Malak Abdullah. 2019. A closer look at fake news detection: A deep learning perspective. In *Proceedings of the 2019 3rd International Conference on Advances in Artificial Intelligence*, pages 24–28.

François Chollet. 2015. keras. https://github.com/fchollet/keras.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Giulia Rizzi Aurora Saibene Berta Chulvi Paolo Rosso Alyssa Lees Jeffrey Sorensen Elisabetta Fersini, Francesca Gasparini. 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.

Dalya Faraj and Malak Abdullah. 2021. Sarcasmdet at semeval-2021 task 7: Detect humor and offensive based on demographic factors using roberta pretrained model. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 527–533.

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.

Vebjørn Isaksen and Björn Gambäck. 2020. Using transfer-based language models to detect hateful and offensive language online. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 16–27.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Casey A Fitzpatrick, Peter Bull, Greg Lipstein, Tony Nelli, Ron Zhu, et al. 2021. The hateful memes challenge: competition report. In *NeurIPS 2020 Competition and Demonstration Track*, pages 344–360. PMLR.

Leonard Williams Levy, Leonard W Levy, Kenneth L Karst, and Adam Winkler. 2000. Encyclopedia of the american constitution.

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*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Vlad Sandulescu. 2020. Detecting hateful memes using a multimodal deep ensemble. *arXiv preprint arXiv:2012.13235*.

Lanyu Shang, Yang Zhang, Yuheng Zha, Yingxi Chen, Christina Youn, and Dong Wang. 2021. Aomd: An analogy-aware approach to offensive meme detection on social media. *Information Processing & Management*, 58(5):102664.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Hajime Watanabe, Mondher Bouazizi, and Tomoaki Ohtsuki. 2018. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE access*, 6:13825–13835.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. https://github.com/facebookresearch/detectron2.