PAIC at SemEval-2022 Task 5: Multi-Modal Misogynous Detection in MEMES with Multi-Task Learning And Multi-model Fusion

Meizhi Jin and **Mengyuan Zhou** and **Mengfei Yuan** and **Dou Hu** and **Xiyang Du** and **Lianxin Jiang** and **Yang Mo** and **XiaoFeng Shi**
Ping An Life Insurance Company of China, Ltd.
{JINMEIZHI005, ZHOUMENGYUAN425, YUANMENGFEI854, HUDOU470, DUXIYANG037, JIANGLIANXIN769, MOYANG853, SHIXIAOFENG309}
@pingan.com.cn

Abstract

This paper describes our system used in the SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification (MAMI) . Multimedia automatic misogyny recognition consists of the identification of misogynous memes, taking advantage of both text and images as sources of information. The task will be organized around two main subtasks: Task A is a binary classification task, which should be identified either as misogynous or not misogynous. Task B is a multi-label classification task, in which the types of misogyny should be identified in potential overlapping categories, such as stereotype, shaming, objectification, and violence. In this paper, we proposed a system based on multi-task learning for multi-modal misogynous detection in memes. Our system combined image features with text features to train a multi-label classification. The prediction results were obtained by the simple weighted average method of the results with different fusion models, and the results of Task A were corrected by Task B. Our system achieves a test accuracy of 0.755 on Task A (ranking 3rd on the final leaderboard) and the accuracy of 0.731 on Task B (ranking 1st on the final leaderboard).

1 Introduction

In the era of mass online communication, more and more people like to share their thoughts on social media platforms. Social media platforms provide users with the ability to express themselves freely.However, it has also led to a rise in cyber hate, such as bullying, sarcasm, and misogyny, on the internet. A study has shown that women have a strong presence online, particularly in image-based social media such as Twitter and Instagram: 78% of women use social media multiple times per day compared to 65% of men. The Web has opened up a whole new world of opportunities for women, but systematic inequality and discrimination in the real world are also found in online spaces in the form of offensive content against them. Now that we live in a world where everything is connected through social media, hate speech against women used to be limited to a specific place or time. There's no need to wait for a specific time or place. All you have to do is type a few keys on the keyboard. Because people on big social media sites make millions of posts every day, it is not possible to manually check all misogynistic posts. To help human curators, it is important to make algorithms that can tell when users post inappropriate content.

Memes are one of the most popular ways to communicate on social media platforms. A meme is essentially an image characterized by a pictorial content with an overlaying text a posteriori introduced by human, with the main goal of being funny and/or ironic. Most of them are made to be funny, but in a short time, people started to use them as a form of hate against women, which led to sexist and aggressive messages in online environments that made the sexual stereotypes and gender inequality in the real world even worse. Misogyny is a type of offensive speech directed at women. It is common on all social media platforms and is becoming more and more of a problem. Previous work on automatic misogyny detection mostly focused on text mode, and they came up with a bunch of supervised methods, like traditional machine learning methods with lexical features and deep learning methods. It can't be certain, though, if memes are misogynistic if only text mode is used to detect them. Memes are intuitively important for automatic misogyny identification tasks.

In this paper, we describe a system that we submitted for detecting misogynous memes. More specifically, we introduce an ensemble model that uses a multi-task learning mechanism based on multi-modal inputs (image information and text information). In our system, image feature representation was extracted by the pre-trained model of ConvNeXt, and text feature representation was ex-

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 555 - 562

tracted by different pre-trained models (deberta-v3-large and roberta-large model). The two features then undergo representation fusion, where they are transformed into reconstructed representation vectors and pumped into a classification layer to yield the final result. Finally, a modality fusion layer performs a weighted average on the fusion results, which results in different text and image features, and the threshold of the final results is adjusted.

2 Background

Many researches worked on identification misogynous in texts. Earlier methods extract carefully engineered discrete features from texts, including n-grams, keyword's sentiment, punctuations, emoticons, etc(Bouazizi and Ohtsuki, 2015), (Ptácek et al., 2014). Recently, researchers have used the powerful technology of deep learning to obtain a more accurate semantic representation of twitter text, including CNN (Convolutional Neural Network), LSTM (Long Short Term Memory) and pre-trained model, etc. Shushkevich and Cardiff (2018) proposed a technique based on combining several simpler classifiers into a more complex blended model. The model considers the probability of classes calculated by simpler models (Logistic Regression, Naive Bayes, and Support Vector Machines - SVM) . Another method from Liu et al. (2018) was proposed, which three classifiers trained by using SVM with a linear kernel, Random Forests (RF) and Gradient Boosted Trees (GBT). In the testing stage, the same way of text pre-processing and feature extraction is applied to test instances separately, and each pair of two out of the three trained classifiers (SVM+RF, SVM+GBT and RF+GBT) are fused by combining the probabilities for each class by averaging.

The primary driver was the renaissance of neural networks, particularly convolutional neural networks (ConvNets). The introduction of AlexNet (Krizhevsky et al., 2017) precipitated the "ImageNet moment", ushering in a new era of computer vision. Since then, this field has developed rapidly. Representative ConvNets like VGGNet (Simonyan and Zisserman, 2015), ResNe(X)t (He et al., 2016), DenseNet (Huang et al., 2017), MobileNet (Howard et al., 2017) and RegNet focused on different aspects of accuracy, efficiency and scalability, and popularized many useful design principles. However, the introduction of Visual Transformer (VIT) completely changed the landscape of network architecture design, which soon replaced ConvNets and became state-of-the-art image classification model. Although the effect of swin transformer is much better than ResNet on the premise of the same size. However, due to the shift operation, it is difficult to design different networks for different sizes of inputs and restart training. And like Detection Transformer (DETR), the convergence is too slow when training. This year, researchers from Facebook AI Research (FAIR) and UC Berkeley reexamined the design space and tested the limits that pure convnet can reach, which shows that the performance of convolutional neural network is no less than that of visual transformer. This series of pure convnet models is named ConvNeXt. It is constructed entirely from standard convnet modules and competes with transformers in terms of accuracy and scalability.

However, little has been revealed by far on how to effectively combine textual and image information to automatic misogyny identification of memes. Schifanella et al. (2016) simply concatenate manually designed features or deep learning based features of texts and images to make prediction with two modalities. Cai et al. (2019) proposed a hierarchical fusion model to deeply fuse three modalities. Different from there works, We proposed a multi-task learning and multi-model.

3 System overview

Figure 1 shows the architecture of our proposed. In this work, we treat text and image classification tasks as two models. Image features are mined using a fine-tuned image classification model based on ConvNeXt. Text features are extracted using a fine-tuned sentence embedding based on the pre-trained RoBERTa model. Image features and text features are fused by direct splicing. A classification layer is connected to the fused model to obtain the category probability of memes. Additionally, we create another fused model that adopts the pre-trained model DeBERTa instead of RoBERTa for text feature exaction progress. Our final submitted labels are voted on by the results inferred by these two proposed models (ConvNeXt + RoBERTa and ConvNeXt + DeBERTa). The threshold value selected for the weighted voting method is 0.405. The classification label returns 1 when the classification probability is greater than 0.405, otherwise it returns 0.

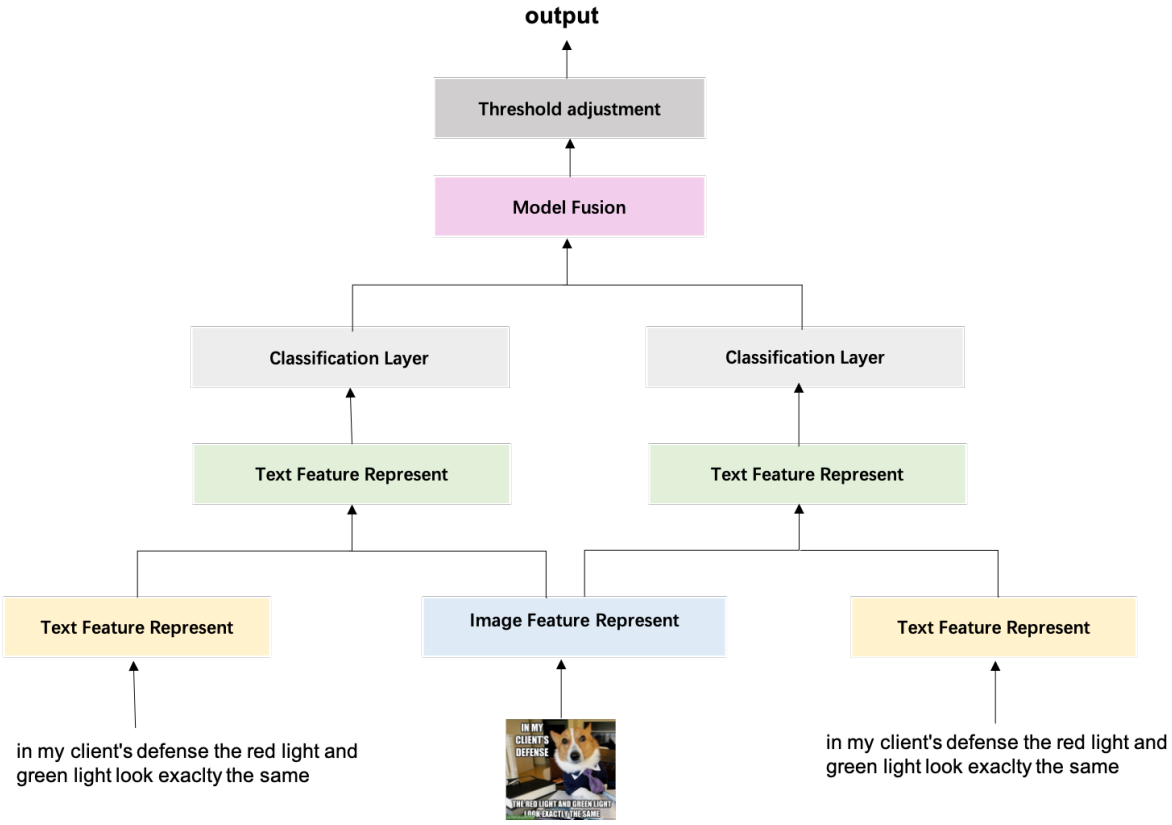

Figure 1: Overview of our proposed model

3.1 Image Feature Representation

In this work, we extracted image feature representation based on an image classification model of ConvNeXt, which trained a baseline model with the Vision Transformer training procedure. Additionally, ConvNeXt (Liu et al., 2022) applied a series of design decisions which are summarized as 1) macro design, 2) ResNeXt, 3) inverted bottleneck, 4) large kernel size, and 5) various layer-wise micro designs. In macro design, ConvNeXt adjusts the number of blocks in each stage and adopts depth-wise convolution. And in micro design, ConvNeXt used fewer activation functions, replacing ReLU with GELU and fewer normalization layers, substituting Batch Normalization (BN) with Layer Normalization (LN), etc. It is worth mentioning that ConvNeXt used data augmentation techniques such as Adamw optimizer, Mixup, Cutmix, RandAugment, Random Erasing, and regularization schemes including Stochastic Depth and Label Smoothing. This was of great help in improving the performance of the model.

3.2 Text Feature Representation

Different pre-trained modals (RoBERTa and DeBERTa) was used as the text feature representation extractor.

RoBERTa - RoBERTa(Liu et al., 2019) is essentially a BERT model with optimal parameters. Compared with BERT model, it used more training data and larger batch size to training longer time. In addition, RoBERTa was trained with dynamic masking instead of static masking, and without NSP loss.

DeBERTa - DeBERTa(He et al., 2021)is a new network architecture proposed by Microsoft. It makes the BERT and RoBERTa models better by using two new techniques. First, the disentangled attention mechanism is applied to represent each word by two vectors, which show their content and relative positions. The attention weights between the word content and position are calculated by disentangled matrices, respectively. Another thing that helps with model pre-training is an "enhanced mask decoder", which adopts the absolute position to predict the masked tokens. In addition, a new virtual adversarial training method is used for fine-tuning to improve models' generalization.

Our system used a series of training techniques. For example, since the general language information and context information obtained by the lower self-attention layer are limited, with the continuous superposition of attention layers, each layer can obtain more language information and context information when encoding. When approaching the last layer, the pre-trained model starts to adjust its embedding information to adapt to different tasks based on different fine-tuning tasks. Therefore, our system set different learning rates in the network to improve the performance of the model. Especially, we set lower learning rates for bottom layers and higher learning rates for top layers. Meanwhile, our system decayed the learning rate with a cosine annealing (Loshchilov and Hutter, 2017) for each batch to improve its overall performance when training deep neural networks.

3.3 Modality Fusion

Multi-modal feature fusion is a useful technique for improving performance in a variety of tasks. According to the order of fusion and prediction, it is divided into early fusion and late fusion. Early fusion is to fuse features before the training model, such as concatenate, add, TFN, MFN, LFN, etc. Late fusion is the fusion of results of different modal predictions, such as maximum fusion, average fusion, etc.

In this scheme, we tried to use different fusion methods for multi-modal fusion, including concatenate, TFN, etc., but finally found that the effect of directly concatenating image features and text features is the best method. Therefore, we concatenate the image features and text features into the fully-connected layer, and then get the fusion features through the normalization layer and multi-sample dropout layer.

3.4 Classification layer

In this work, our system used a two-layer fully-connected neural network as our classification layer. The activation functions for the hidden layer and the output layer are the ReLu and Sigmoid functions, respectively. The loss function is a BCE loss. In addition, we also tried to use SVM as the classification layer, but the performance was not effective.

3.5 Model Fusion

Our system used different pre-trained models to extract text features (RoBERTa and DeBERTa). After fusing image features, we trained a multi-label classification model to obtain the probability value of each category (misogyny, stereotype, shading, objective, and violence). We used weighted voting to fuse the results with two fusion models (ConvNeXt + RoBERTa, and ConvNeXt + DeBERTa). The threshold value selected for the weighted voting method is 0.405 (the classification label returns 1 when the classification probability is greater than 0.405, otherwise it returns 0). Through analysis, it is found that when "shaming", "stereotype", "objectification", and "violence" are positive samples, the "misogynous" must also be positive. In order to improve the results of Task A, the prediction results of "misogynous" are corrected by the prediction results of "shaming", "stereotype", "objectification" and "violence". Experiments show that it is a significant improvement over the result of Task A. Moreover, we also tried to use the results of "shaming", "stereotype", "objectification" and "violence" to correct the results of "misogynous", but found the effect was not satisfying.

4 Experimental setup

4.1 Dataset

The datasets for the MAMI competition(Fersini et al., 2022) are memes collected from the web and manually annotated via crowd sourcing platforms. As summarized in Table 1, the organizers provided 10,000 memes (in jpg format) and a csv file for the training set and 1,000 memes (in jpg format) and a csv file for the testing set. For both Subtask A and B, the memes are released as jpg images.

misogynous: a binary value (1/0) indicating if the meme is misogynous or not. A meme is misogynous if it conceptually describes an offensive, sexist or hateful scene (weak or strong, implicitly or explicitly) having as target a woman or a group of women.

shaming: a binary value (1/0) indicating if the meme is denoting shaming. A shaming meme aims at insulting and offending women because of some characteristics of the body.

stereotype: a binary value (1/0) indicating if the meme is denoting stereotype. A stereotyping meme aims at representing a fixed idea or set of (physically or ideologically) characteristics of women.

objectification: a binary value (1/0) indicating if the meme is denoting objectification. A meme that describes objectification represents a woman like an object through over-appreciation of physical

Task A	Training	Testing	Task B	Training	Testing
Misogynous	5000	500	stereotype	1271	146
			shaming	2810	350
			objectification	2201	348
			violence	953	153
No Misogynous	5000	500	stereotype	3	0
			shaming	0	0
			objectification	1	0
			violence	0	0
Total	10000	1000	Total	7239	997

Table 1: Dataset label distribution

appeal (sexual objectification) or depicting woman as an object (human being without any value as a person).

violence: a binary value (1/0) indicating if the meme is denoting violence. A violent meme describes physical or verbal violence represented by textual or visual content.

The label distribution related to the training and testing datasets is reported in Table1. While the distribution of labels related to the field of misogynous is balanced (for both testing and training datasets), the classes related to the other fields are quite unbalanced. Furthermore, when memes are labeled as "No Misogynous," the classes of "shaming," "stereotype," "objectification," and "violence" are found to be positive in testing datasets. However, there is some error data in training datasets where memes are identified as "No Misogynous".

4.2 Training Details

Image datasets. The public pre-trained model of ConvNeXt_base_22k_1k_384 is adopted in the proposed model. Preprocess the image data through random horizontal and vertical clipping, Random Affine, ColorJitter, Normalize, etc. It is worth mentioning that we set the image size of the training datasets to 384 x 384 and the image size of the verification and testing datasets to (384 + 32) x (384 + 32), which can improve the generalization of the model.

Text datasets. The pre-trained model of deberta-v3-large and roberta-large are adopted for analyzing the text. Models download from HuggingFace[1] are directly used in our model.

Learning rate initialization. Our system sets different learning rates for each layer of text pre-trained model. In particular, the layers of the pre-trained model are divided into three groups with distinct hyper parameters. The learning rates for layer-0 to layer-7 were set to 1e-5/2.6, while layer-8 to layer-15 were set to 1e-5, and layer-16 to layer-23 were set to 1e-5 * 2.6. Setting different learning rates for different layers can make the training more effective and improve the performance of the model.

Learning rate decay. Our system used a method of cosine annealing to decay the learning rates for each batch to avoid falling into a local optimal solution.

multi sample dropout. The four samples dropout were used in our system after fusing the image features and text features. Experiments show that the performance and effects have been significantly improved.

Optimization. Our system applied the AdamW optimizer to optimize the loss function.

Loss Function. Our system applied the BCE loss for multi-label binary classification.

Evaluation Function. Due to our mistakes, we used the incorrect evaluation criteria, which used macro-average F1-Measure instead of weighted-average F1-Measure to train the multi-label classification model. Through offline experiments, we discovered that using the correct evaluation criteria training can result in better and more stable results.

5 Results

In this section, the results of the experiments for our runs will be discussed and compared to the results published in Table 2. We try to only use image or text features for training through different pre-trained models, including Swin Transformers, ConvNeXt, RoBERTa, and DeBERTa. In addition, we also used some other training techniques, such as training a multi-label classification model with

[1] https://huggingface.co/models

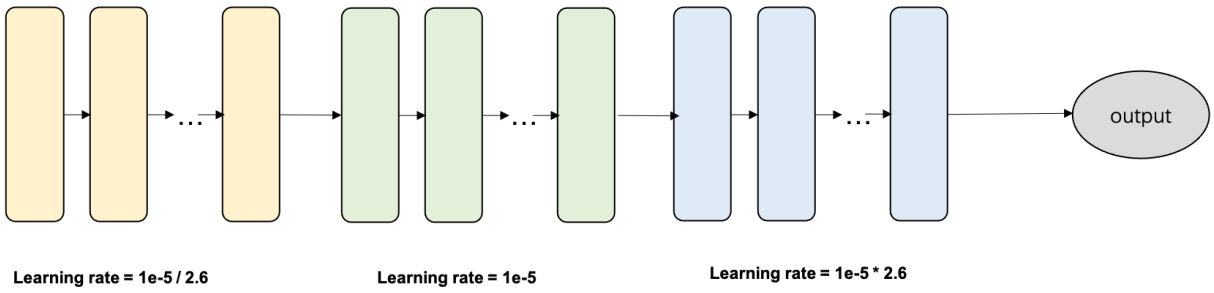

Figure 2: Learning rate initialization

five or four classes, using different classification layers (SVM or full-connected layer), using different fusion methods (concat or LFN), and adjusting the threshold value of fusion prediction results, etc. The specific result of the analysis is shown in Table 2.

V1: Training a image classification model based on a fine-tuned image classification model grounded on swin_large_patch4_window12_384_22k.

V2: Training an image classification model based on a fine-tuned image classification model grounded on ConvNeXt_base_22k_1k_384.

V3: Training a text classification model based on a pre-trained model of roberta-large.

V4: Training a text classification model based on a pre-trained model of deberta-v3-large.

V5: Training a binary classification model for Task A and a multi-label classification model for Task B using four classes (stereotype, humiliation, objectification and violence) after fusing the image features by ConvNeXt_base_22k_1k_384 and the text features by roberta-large, respectively.

V6: Training a multi-label classification model that uses four classes (stereotype, humiliation, objectification and violence) after fusing the image features by ConvNeXt_base_22k_1k_384 and text features by roberta-large (the misogynous results are obtained by the multi-label predictions).

V7: The only difference between this scheme and V6 is that LFN uses feature fusion rather than simply concatenating.

V8: This scheme is the same as the V6. The only difference is that the training labels contain misogynous content (five classes of misogynous, stereotype, humiliation, objectification, and violence).

V9: This scheme is the same as V6. The only difference is that the text features extractor was replaced by deberta-v3-large.

V10: In this scheme, the results of V8 and V9 are fused by a weighted average approach with a threshold set to 0.5.

V11: This scheme is the same as in v10. The only difference is that the threshold is adjusted to 0.4.

V12: This scheme is the same as in v10. The only difference is that the threshold is adjusted to 0.405 (Final results of the leaderboard).

V13: This scheme is the same as in v10. The only difference is that the threshold is adjusted to 0.403 (results not submitted).

Through the experiment, we found the following conclusions:

- The effect of ConvNeXt_base_22k_1k_384 is better than swin_large_patch4_window12_384_22k in our system.
- The effect of multi-modal (text and image) is better than single-modal (text or image) in our system.
- In our system, the effect of multi-task learning is better than individual training.
- According to the results of stereotype, humiliation, objectification and violence to correct misogynous results have a significant improvement effect in our system.
- Adjusting the prediction results also has an improvement for Task B in our system.
- The simple concatenate method is better than the complex feature fusion method in our system.

Version A	Model	Task A	Task B
V1	$swin\_large\_patch4\_window12\_384\_22k$	0.6135	0.5965
V2	$ConvNeXt\_base\_22k\_1k\_384$	0.6431	0.6229
V3	roberta-large	0.656	0.6389
V4	deberta-v3-large	0.6947	0.649
V5	ConvNeXt + RoBERTa + TaskA + TaskB + 4 categories	0.7053	0.7136
V6	ConvNeXt + RoBERTa + TaskB/A + 4 categories	0.7514	0.7136
V7	ConvNeXt + RoBERTa + TaskB/A + 4 categories + LFN	0.7464	0.704
V8	ConvNeXt + RoBERTa + TaskB/A + 5 categories	0.7643	0.7129
V9	ConvNeXt + DeBERTa + TaskB/A + 5 categories	0.7727	0.7240
V10	v8+v9+0.5	0.7785	0.719
V11	v8+v9+0.4	0.7554	0.728
V12	v8+v9+0.405	0.7566	0.7307
V13	v8+v9+0.403	0.7569	0.7319

Table 2: Experimental results of task a and Task B

6 Conclusion

In this paper, we proposed a system to make full use of two modes (image and text) for solving the challenging multi-mode misogynous meme detection task. We extracted the image features and text features using different pre-trained models and weighted averaged the results after fusing the features of image and text. The system performs well in identifying misogynistic memes, achieving 77.85% percent accuracy on TaskA and 73.1% percent accuracy on TaskB.

7 Acknowledgments

This research was supported by the PingAn Life Insurance. We thank the organizers of Università degli Studi di Milano Bicocca (Italy), Google Jigsaw (New York) and Universitat Politècnica de València (Spain) for their support.

References

Mondher Bouazizi and Tomoaki Ohtsuki. 2015. Sarcasm detection in twitter: "all your products are incredibly amazing!!!" - are they really? In *2015 IEEE Global Communications Conference, GLOBECOM 2015, San Diego, CA, USA, December 6-10, 2015*, pages 1–6. IEEE.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2506–2515. Association for Computational Linguistics.

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.

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 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2261–2269. IEEE Computer Society.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90.

Han Liu, Fatima Chiroma, and Mihaela Cocea. 2018. Identification and classification of misogynous tweets using multi-classifier fusion. In *Proceedings of the Third Workshop on Evaluation of Human Language*

Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018, volume 2150 of *CEUR Workshop Proceedings*, pages 268–273. CEUR-WS.org.

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. *CoRR*, abs/1907.11692.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. *CoRR*, abs/2201.03545.

Ilya Loshchilov and Frank Hutter. 2017. SGDR: stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Tomás Ptácek, Ivan Habernal, and Jun Hong. 2014. Sarcasm detection on czech and english twitter. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 213–223. ACL.

Rossano Schifanella, Paloma de Juan, Joel R. Tetreault, and Liangliang Cao. 2016. Detecting sarcasm in multimodal social platforms. In *Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016, Amsterdam, The Netherlands, October 15-19, 2016*, pages 1136–1145. ACM.

Elena Shushkevich and John Cardiff. 2018. Classifying misogynistic tweets using a blended model: The AMI shared task in IBEREVAL 2018. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, volume 2150 of *CEUR Workshop Proceedings*, pages 255–259. CEUR-WS.org.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.