AIDA-UPM at SemEval-2022 Task 5: Exploring Multimodal Late Information Fusion for Multimedia Automatic Misogyny Identification

Álvaro Huertas-García
Universidad Politecnica de Madrid
Madrid, Spain
Universidad Rey Juan Carlos de Madrid
Madrid, Spain
alvaro.huertas.garcia@alumnos.upm.es

Helena Liz López
Universidad Politecnica de Madrid
Madrid, Spain
Universidad Rey Juan Carlos de Madrid
Madrid, Spain
helena.liz@alumnos.upm.es

Guillermo Villar-Rodríguez
Universidad Politecnica de Madrid
Madrid, Spain
guillermo.villar@upm.es

Alejandro Martín
Universidad Politecnica de Madrid
Madrid, Spain
alejandro.martin@upm.es

Javier Huertas-Tato
Universidad Politecnica de Madrid
Madrid, Spain
javier.huertas.tato@upm.es

David Camacho
Universidad Politecnica de Madrid
Madrid, Spain
david.camacho@upm.es

Abstract

This paper describes the multimodal late fusion model proposed in the SemEval-2022 Multimedia Automatic Misogyny Identification (MAMI) task. The main contribution of this paper is the exploration of different late fusion methods to boost the performance of the combination based on the Transformer-based model and Convolutional Neural Networks (CNNs) for text and image, respectively. Additionally, our findings contribute to a better understanding of the effects of different image preprocessing methods for meme classification. We achieve 0.636 F1-macro average score for the binary sub-task A, and 0.632 F1-macro average score for the multi-label sub-task B. The present findings might help solve the inequality and discrimination women suffer on social media platforms.

1 Introduction

The proposed task, SemEval-2022 Task 5 Multimedia Automatic Misogyny Identification (MAMI) (Fersini et al., 2022) consists in the identification of misogynous memes in English language, taking advantage of both text and images available as a source of information.

Overall, our proposed method consists of a multimodal approach combining different features (e.g., logits, probabilities, embeddings) of a text Transformer-based model and an image CNN model in a late fusion approach. This late fusion step implies that both models are trained and fine-tuned separately to the task. Then, the features from each model are concatenated and jointly used as input for a final classifier that combines their knowledge to obtain a final prediction (see Figure 1). Different preprocessing steps, text and image models, concatenated feature combinations, and classifiers are explored to obtain the final multimodal architecture.

Our presented method has been developed for sub-task A and B from the MAMI competition independently. sub-task A consists of misogynous meme binary classification, where a meme should be categorized either as misogynous or not misogynous. On the other hand, sub-task B requires a more detailed multi-label classification where misogynous content should be recognized among potential overlapping categories such as stereotype, shaming, objectification, and violence.

It is noteworthy that our multimodal late fusion method outperforms single models in both sub-tasks, being more remarkable in complex sub-task B. Similarly, considering that both sub-tasks share the same data, the results of model evaluation on both tasks show how the model trained on complex sub-task B can achieve the same results as a model trained only on binary sub-task A. Therefore, future studies should investigate the complexity and pruning of the required models.

This paper provides new insights into information fusion for tackling multimodal tasks, presenting an in-depth exploration of different late fusion approaches and image processing steps. The presented work might help identify malicious be-

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 771 - 779

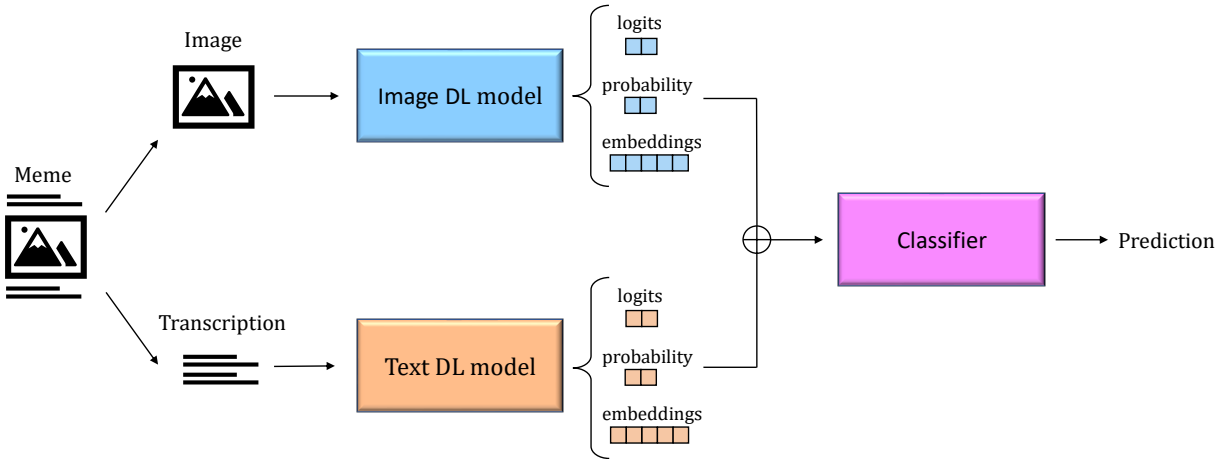

Figure 1: Summarized diagram of the late fusion multimodal system proposed for misogyny detection

haviours towards women on social media.

2 Background

Misogyny comprises every hateful and prejudicial action against women, ranging from discrimination, objectification, violence and disdain affecting women to all the types of manifested male superiority like patriarchy, androcentrism and privilege (Pamungkas et al., 2020). Misogynist posts represent one type of hate speech on Online Social Networks, but it is complex to distinguish them from other sorts of offensive discourses in an automated way (Shushkevich and Cardiff, 2019).

Whereas research shows a survey of methods for misogyny recognition in text including traditional methods and their ensembles and neural networks (Shushkevich and Cardiff, 2019), studies reviewing mysogyny in images are more scarce, and more work can may be found under the generalization of sexist content (Campisi et al., 2018; Fersini et al., 2019) than under the particular topic of misogyny.

These sexist images can be in the form of memes, multimedia content with a humorist goal composed of photos and/or illustrations with some text on it (Sabat et al., 2019). Furthermore, this initially shows that sexist and also misogynist content on social networks do not need just the isolated use of image and text processing but the combination of both. In fact, having images or texts on their own may not lead to hateful speech and only their mix is offensive (Sabat et al., 2019), underlining the major necessity of architectures that encode the global meaning of memes.

On the one hand, the latest advances in NLP for text classification include the use of Transformers, which has demonstrated to be successful in the detection of misogyny (Samghabadi et al., 2020; Aldana-Bobadilla et al., 2021). These distributional models encode the meaning of texts in vectors that also capture the context (Devlin et al., 2018). However, they may not be so optimal for the detection of subclasses, for instance between the absence and the implicit or sarcastic presence of aggressiveness (Samghabadi et al., 2020), or when the female subject that is attacked is not present in the text (Aldana-Bobadilla et al., 2021).

On the other hand, when images are also taken into account, the use of Convolutional Neural Networks (CNNs) for images is still present (Gomez et al., 2020) through image-based state-of-the-art models such as `VGG16`, `ResNet`, `DenseNet` and `Inception`. Specifically, `VGG16` works by itself for the prediction of offensive memes and succeeds when combined with different models for the text inside them (LSTMs, BiLSTMs and CNNs), which were later compared (Aman et al., 2021), but when the comparison is among vision models, depending on the dataset and the multimodal approach followed, `ResNet50` can surpass `VGG16` and `ResNet152` for VGCN-BERT combined pipelines for hateful images detection (Vlad et al., 2020), but `Inception` outperforms the F1-scores from `ResNet50` for multimodal approaches with BiLSTMs for 'troll' (sarcastic or offensive) meme detection (Hossain et al., 2021), and just using `DenseNet` alone for meme emotion recognition can be even better than `ResNet` alone and than either `DenseNet` or `ResNet` in image joining forces with BERT for the textual features (Guo et al., 2020). Alternative implementations for vision are Transformers such as VisualBERT, which represents an architecture with image and text mod-

els (Lippe et al., 2020) or CLIP, which predicts the text that better describes the images (Zia et al., 2021).

Recent research has shown that joining the probabilities from the CNN classification and the output statistical variables obtained after training a successive classifier (Huertas-Tato et al., 2022) improves the task. These results invite to use this advance in information fusion for practical applications such as hateful images and, specifically, misogynist memes. In line with this theoretical framework, our work will concatenate the different outputs (e.g, logits, probabilities, last hidden embeddings) from CNN-based image classification and from Transformers-based text classification towards a better performance in MAMI.

3 System overview

As previously mentioned, our proposed approach is depicted in Figure 1. This section details the preprocessing steps, the text and image models employed, and the methodology followed to train the final multimodal late fusion classifier that combines different features.

3.1 Image Preprocessing

To rule out the possibility of text misleading the image CNN model and to enhance its focus on the image, three different preprocessing steps are separately explored. Consequently, three different image models are trained: (1) no preprocessing, (2) blacking out, and (3) inpainting the text from the image. These preprocessing methods make use of EasyOCR[1] and OpenCV (Bradski, 2000). Figure 2 illustrate some examples of the results of these preprocessing steps. Additionally, an application for applying the inpainting preprocessing to images has been publicly published with the aim of contributing to the scientific community[2].

3.2 Text Preprocessing

Ftfy package (Speer, 2019) was used as a preprocessing step for fixing text and ensuring it is uniformly UTF-8 encoded. URLs, emojis, or other native features present in the text are not modified as we consider these characteristics crucial for this task.

[1] https://github.com/JaidedAI/EasyOCR

[2] https://huggingface.co/spaces/Huertas97/Inpaint_Me

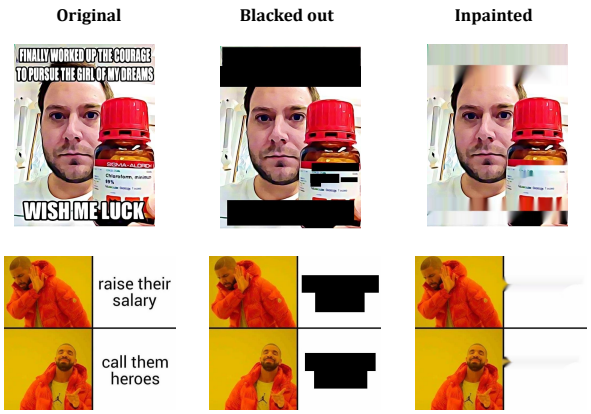

Figure 2: Example of different image preprocessing steps

3.3 Explored Models

According to the models employed, it is worth mentioning that different Transformer-based models publicly available at Hugging Face (Wolf et al., 2020) are evaluated as the textual model:

- `bertweet-base` (Nguyen et al., 2020): large-scale language model pre-trained for English Tweets based on RoBERTa (Liu et al., 2019) pre-training procedure and BERT architecture (Devlin et al., 2019).
- `all-distilroberta-v1`[3]: pre-trained `distilroberta-base` (Sanh et al., 2019) model fine-tuned on a 1B sentence pairs dataset using a contrastive learning objective.
- `all-miniLM-L6-v2`[4]: pre-trained Microsoft `MiniLM` (Wang et al., 2020) model fine-tuned on a 1B sentence pairs dataset using a contrastive learning objective.
- `twitter-roberta-base-offensive` (Barbieri et al., 2020): `roberta-base` model trained on 58M tweets and finetuned for offensive language identification with the TweetEval benchmark.
- `twitter-xlm-roberta-base` (Barbieri et al., 2021): `xlm-roberta-base` model trained on 198M multilingual tweets.

For image processing we have used CNNs, which extract features with convolutional layers and deduce knowledge with dense layer. We use

[3] https://huggingface.co/sentence-transformers/all-distilroberta-v1

[4] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Optimization	Hyperparameters	Values
Text Model	learning rate	min = 1e-6 , max = 1e-3
	epochs	min = 1, max = 10
	weight decay	min = 0 , max = 1
	gradient accumulation steps	min = 1 , max = 4
	scheduler	constant_schedule constant_schedule_with_warmup linear_schedule_with_warmup cosine_schedule_with_warmup cosine_with_hard_restarts_schedule_with_warmup polynomial_decay_schedule_with_warmup
	optimizer	AdamW
	sliding window	True, False
	pos_weigths*	[2, 1, 1, 2], [2, 0.5, 0.5, 2], [1, 1, 1, 1]
Image Data augmentation	shear range	0.1
	zoom range	0.1
	rotation range	45
	width shift range	0.1
	height shift range	0.1
	horizontal flip	True
	brightness range	0.7-1.1
	channel shift range	0.05
Image Model	optimizer	Adam
	learning rate	0,001
	preprocess_input	True
	pos_weigths*	[1,1,1,1], [1, 3.96, 1.78, 2.27, 5.25]
	percentage of frozen layers	0.1, 0.3*
Auto-sklearn	time_left_for_this_task	60, 120, 500, 3600, 7200
	memory_limit	6072
	exclude	None
	resampling_strategy	Cross Validation 5 folds
	ensemble_size	10

Table 1: Hyperparameters optimized during the development of the proposed approaches. *Only for multi-label subtask B.

different architectures pretrained from state of the art:

- `VGG16` (Simonyan and Zisserman, 2014): it is composed by a feed-forward set of units and is the most straightforward without additional forward connections or auxiliary outputs. However this architecture has to adjust lots of parameters.

- `ResNet50` (He et al., 2016): The main advantage of this architecture is the shortcut connections, these links skip one or more layers, aggregating their output to the outputs of the stacked layers.

- `DenseNet201` (Huang et al., 2017): Instead of adding more layers to the architecture, it increases the number of connections between units, connecting every units with later ones.

- `Inception_v3` (Szegedy et al., 2016): it factorises the convolution into smaller ones (that can be asymmetric) to reduce the cost. Moreover, this architecture has an auxiliary classifier between layers, that acts as regularizer.

- `EfficientNetB0` (Tan and Le, 2019): This architecture uses a compound coefficient to scale all dimensions of depth, width and resolution.

Finally, the Auto-sklearn package (Feurer et al., 2015, 2020) is used to automatically explore a wide range of models and preprocessing approaches available in scikit-learn and identify the best ensemble configuration for the multimodal late fusion step. We opted for this method because it implements Bayesian Optimization for searching the optimal pipeline configuration and Ensemble Selection to choose the suitable model.

3.4 Multimodal late fusion approaches

From the text and image models, three features are used for the late fusion step; the output of the activation function from the last classification layer (i.e., probabilities) and its input (i.e., logits), and the vectorize output representation of the last hidden layer (i.e., embeddings). In order to develop the final multimodal classification predictions, different late fusion approaches for combining these features are considered.

Naive baseline approaches consist of averaging or taking the maximum logit or probability values from both models for each class or label depending on the sub-task. An advanced baseline approach consists of finding the weights for each model that will give the lowest mean square error (MSE) score between multimodal predictions and real values using logits or probabilities. For this purpose, Sequential Least Squares Programming (SQLSP) from Scipy package (Virtanen et al., 2020) is the optimization method used. Finally, logits, probabilities or embeddings from both models are used as input in Bayesian Optimization for searching the optimal pipeline configuration and Ensemble Selection using Auto-sklearn.

4 Experimental Setup

As previously mentioned, both sub-tasks share the same data and the official metric for system evaluation, F1-macro averaged. The dataset consists of 10.000 memes and its corresponding text transcriptions. To develop the proposed approach, balanced data for sub-task A is split into 64% train, 16% validation and 20% test in a stratified way using scikit-learn package (Pedregosa et al., 2011) with 42 as random state. Regarding multi-label sub-task B where the data is unbalanced, the data is split into 64% train, 16% validation and 20% test using the "iterative_train_test_split" method from scikit-multilearn package (Szymański and Kajdanowicz, 2018) to equally represent the different combination of overlapping labels in the splits.

To obtain the best results and avoid overfitting, we optimized several hyperparameters. Table 1 summarizes the hyperparameters tuned for both sub-tasks using their respective development sets. The experiment tracking and the selected hyperparameter values are published in Weight and Biases[5][6]. The resulting model are openly available in HuggingFace[7].

5 Results

5.1 Sub-task A - Binary misogyny classification

5.1.1 Image

Firstly, we analysed which of the three preprocessing techniques performed best for this task, where we observed that the images without preprocessing showed the best results. Therefore, all models were trained on this dataset. Table 2 shows performance from the five models, where the best model is `EfficientNetB0` which achieves the highest F1 score.

Image Model	F1 macro Avg
VGG16	0.5224
ResNet	0.6143
DenseNet	0.6608
EfficientNet	**0.6825**
Inception	0.6792

Table 2: Evaluation results of image models in the validation split of subtask A.

5.1.2 Text

The Trasnforme-based models results for validation split are shown in Table 3. As can be derived from these results, `bertweet-base` model has the best score and it is the one selected for the next multimodal late fusion step.

5.1.3 Multimodal Late Fusion

As explained in 3.4 different late fusion methods are explored. The best score is obtained using Auto-sklearn and probabilities from both text and image models as input data (see Table 4). The best Auto-sklearn ensemble configuration is composed

[5]Tracking experiments W&B sub-task A
[6]Tracking experiments W&B sub-task B
[7]Multi-label sub-task B model in Hugging Face hub

Text Model	F1 macro Avg
bertweet-base	**0.8320**
all-miniLM-L6-v2	0.8254
all-distilroberta-v1	0.8239
twitter-roberta-base-offensive	0.8082
twitter-xlm-roberta-base	0.7950

Table 3: Evaluation results of Transformer-based models in the validation split of subtask A.

Late Fusion Method	F1 macro Avg
Avg Logit	0.7874
Avg Probs	0.8410
Max Logit	0.8425
Max Probs	0.8410
Weighted Avg Logit	0.8307
Weighted Avg Probs	0.8430
Auto-sklearn Logit	0.8400
Auto-sklearn Probs	**0.8430**
Auto-sklearn Embs	0.7890

Table 4: Evaluation results of multimodal late fusion methods in the validation split of subtask A.

of three SGD classifiers [8].

These scores presented are remarkable, as logits contain more information about the model's decisions, but the concatenation of probabilities as late fusion input proves to be more useful for sub-task A. This might be explained by the fact that logits from different models have different distributions, not being as useful as normalized inputs.

5.2 Sub-task B - Multi-label misogyny classification

5.2.1 Image

In multilabel task the preprocessing is interesting. The third technique, inpainting the text from the images has better performance than the no preprocessing, however, the difference between them is low (as f1 score is 0.02) so it was decided to continue with the non-preprocessed images in order to maintain the methodology of the sub-task A. In this sub-task the best model is `EfficientNetB0`, as in the first one, which reached the highest performance.

[8] https://github.com/AIDA-UPM/AIDA-UPM-SemEval-2022-Task-5-MAMI-

Image Model	F1 macro Avg
VGG16	0.2626
ResNet	0.27734
DenseNet	0,2857
EfficientNet	**0.3477**

Table 5: Evaluation results of image models in the validation split of subtask B.

5.2.2 Text

As in sub-task A, `bertweet-base` model has the best score and it is the one selected for the next multimodal late fusion step (see Table 6).

Text Model	F1 macro Avg
bertweet-base	**0.5785**
all-distilroberta-v1	0.5570
twitter-roberta-base-offensive	0.4666
twitter-xlm-roberta-base	0.4218
all-miniLM-L6-v2	0.2057

Table 6: Evaluation results of Transformer-based models in the validation split of subtask B.

5.2.3 Multimodal Late Fusion

As in sub-task A, Auto-sklearn and probabilities from both text and image models as input data (see Table 7) shows the best results. The Auto-sklearn ensemble configuration is composed of Random Forest MLP, and Naive Bayes classifiers.

It is interesting to note that in a more in-depth analysis of the classification results performed by the different fusion methods, the simplest (e.g., average, max) only learned to correctly separate the majority label (misogynous or non-misogynous). However, the models using Auto-sklearn did manage to also classify the less frequent labels.

Finally, we report our test competition results along with the baseline results from the organizers of the competition. For sub-task A, the baselines are grounded a fine-tuned sentence embedding using the USE pre-trained model; fine-tuned image classification model grounded on `VGG-16`; and a concatenation of deep image and text representations using a single layer neural network. For subtask B, the baselines are grounded on a multi-label model, based on the concatenation of deep image and text representations; a hierarchical multi-label model, based on text representations, for predicting if a meme is misogynous or not and, if misogynous, the corresponding type.

Late Fusion Method	F1 macro Avg
Avg Logit	0.4475
Avg Probs	0.2014
Max Logit	0.4289
Max Probs	0.3308
Weighted Avg Logit	0.4977
Weighted Avg Probs	0.1627
Auto-sklearn Logit	0.5411
Auto-sklearn Probs	**0.5522**
Auto-sklearn Embs	0.3897

Table 7: Evaluation results of multimodal late fusion methods in the validation split of subtask B.

Subtask	Method	F1 macro Avg
A	Our Late Fusion method	0.636
	Baseline_Text	0.640
	Baseline_Image	0.639
	Baseline_Image_Text	0.543
B	Our Late Fusion method	0.632
	Baseline_Hierarchical_M.	0.621
	Baseline_Flat_Multilabel	0.421
	Baseline_Image_Text	0.000
	Baseline_Text	0.000
	Baseline_Image	0.000

Table 8: Competition results

6 Conclusion

As a conclusion of the results obtained in the exploration and evaluation of models for the development of the multimodal late information fusion architecture, it is evident that the contribution between image and text is different, being the text much more informative in both sub-tasks.

In the case of the image model, it is interesting to note that the different proposed pre-processing techniques do not seem to have a beneficial effect on model training. Although further future analysis is needed, one possible justification could be that the information supplied by the text present in the images provides valuable information rather than noise in the CNN models.

Finally, it is also important to point out that in sub-task A, the multimodal strategy is not as relevant as expected as the baseline strategy that combines image and text information and our approach has the lowest scores in Table 8. A possible explanation for the results obtained could be the difference distribution between train and test sets of the competition, which would have facilitated overfitting in the development of the models in sub-task A. On the contrary, our method proves to be beneficial in sub-task B. Therefore, this could reinforce the idea of overfitting in sub-task A since sub-task B is more complex and better results are obtained. Following the same line, obtaining text models in sub-task B that maintain sub-task A performance supports future research exploring pruning techniques to avoid this situation. Additionally, this study provides a springboard for exploring the late fusion methods applied in this work in different modal problems and other domain scenarios, and comparing them to an end-to-end deep classifier.

In general, our results from the presented multimodal late fusion approach are encouraging to counteract malicious misogynistic behavior against women on social media.

Acknowledgments

This work has been supported by the research project CIVIC: Intelligent characterisation of the veracity of the information related to COVID-19, granted by BBVA FOUNDATION GRANTS FOR SCIENTIFIC RESEARCH TEAMS SARS-CoV-2 and COVID-19, by the Spanish Ministry of Science and Innovation under FightDIS (PID2020-117263GB-100) and XAI-Disinfodemics (PLEC2021-007681) grants, by Co-

munidad Autónoma de Madrid under S2018/TCS-4566 grant, by European Comission under IBERIFIER - Iberian Digital Media Research and Fact-Checking Hub (2020-EU-IA-0252), by "Convenio Plurianual with the Universidad Politécnica de Madrid in the actuation line of *Programa de Excelencia para el Profesorado Universitario*" and by the research project DisTrack: Tracking disinformation in Online Social Networks through Deep Natural Language Processing, granted by Barcelona Mobile World Capital Foundation.

References

Edwin Aldana-Bobadilla, Alejandro Molina-Villegas, Yuridia Montelongo-Padilla, Ivan Lopez-Arevalo, and Oscar S Sordia. 2021. A language model for misogyny detection in latin american spanish driven by multisource feature extraction and transformers. *Applied Sciences*, 11(21):10467.

Aayush Aman, Gopal Krishna, Tushar Anand, and Anubhaw Lal. 2021. Identification of offensive content in memes. In *Data Science and Security*, pages 438–445. Springer.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2021. Xlm-t: A multilingual language model toolkit for twitter.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

G. Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Gaia Campisi, Silvia Corchs, Elisabetta Fersini, Francesca Gasparini, and Monica Mantovani. 2018. Automatic detection of sexist content in memes. *Image*, 46:53–9.

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*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Elisabetta Fersini, Francesca Gasparini, and Silvia Corchs. 2019. Detecting sexist meme on the web: A study on textual and visual cues. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 226–231. IEEE.

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.

Matthias Feurer, Katharina Eggensperger, Stefan Falkner, Marius Lindauer, and Frank Hutter. 2020. Auto-sklearn 2.0: Hands-free automl via meta-learning. *arXiv:2007.04074 [cs.LG]*.

Matthias Feurer, Aaron Klein, Jost Eggensperger, Katharina Springenberg, Manuel Blum, and Frank Hutter. 2015. Efficient and robust automated machine learning. In *Advances in Neural Information Processing Systems 28 (2015)*, pages 2962–2970.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1470–1478.

Xiaoyu Guo, Jing Ma, and Arkaitz Zubiaga. 2020. Nuaa-qmul at semeval-2020 task 8: Utilizing bert and densenet for internet meme emotion analysis. *arXiv preprint arXiv:2011.02788*.

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.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2021. Nlp-cuet@ dravidianlangtech-eacl2021: Investigating visual and textual features to identify trolls from multimodal social media memes. *arXiv preprint arXiv:2103.00466*.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

Javier Huertas-Tato, Alejandro Martín, Julian Fierrez, and David Camacho. 2022. Fusing cnns and statistical indicators to improve image classification. *Information Fusion*, 79:174–187.

Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, and Helen Yannakoudakis. 2020. A multimodal framework for the detection of hateful memes. *arXiv preprint arXiv:2012.12871*.

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*.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. Misogyny detection in twitter: a multilingual and cross-domain study. *Information Processing & Management*, 57(6):102360.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Benet Oriol Sabat, Cristian Canton Ferrer, and Xavier Giro-i Nieto. 2019. Hate speech in pixels: Detection of offensive memes towards automatic moderation. *arXiv preprint arXiv:1910.02334*.

Niloofar Safi Samghabadi, Parth Patwa, Srinivas Pykl, Prerana Mukherjee, Amitava Das, and Thamar Solorio. 2020. Aggression and misogyny detection using bert: A multi-task approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 126–131.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Elena Shushkevich and John Cardiff. 2019. Automatic misogyny detection in social media: A survey. *Computación y Sistemas*, 23(4):1159–1164.

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

Robyn Speer. 2019. ftfy. Zenodo. Version 5.5.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Piotr Szymański and Tomasz Kajdanowicz. 2018. A scikit-based python environment for performing multi-label classification.

Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

George-Alexandru Vlad, George-Eduard Zaharia, Dumitru-Clementin Cercel, and Mihai Dascalu. 2020. Upb@ dankmemes: Italian memes analysis-employing visual models and graph convolutional networks for meme identification and hate speech detection. *EVALITA Evaluation of NLP and Speech Tools for Italian-December 17th, 2020*, page 288.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Haris Bin Zia, Ignacio Castro, and Gareth Tyson. 2021. Racist or sexist meme? classifying memes beyond hateful. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 215–219.