UPB at SemEval-2022 Task 5: Enhancing UNITER with Image Sentiment and Graph Convolutional Networks for Multimedia Automatic Misogyny Identification

Andrei Paraschiv, Mihai Dascalu, Dumitru-Clementin Cercel
University Politehnica of Bucharest, Faculty of Automatic Control and Computers
{andrei.paraschiv74, mihai.dascalu, dumitru.cercel}@upb.ro

Abstract

In recent times, the detection of hate-speech, offensive, or abusive language in online media has become an important topic in NLP research due to the exponential growth of social media and the propagation of such messages, as well as their impact. Misogyny detection, even though it plays an important part in hate-speech detection, has not received the same attention. In this paper, we describe our classification systems submitted to the SemEval-2022 Task 5: MAMI - Multimedia Automatic Misogyny Identification. The shared task aimed to identify misogynous content in a multi-modal setting by analysing meme images together with their textual captions. To this end, we propose two models based on the pre-trained UNITER model, one enhanced with an image sentiment classifier, whereas the second leverages a Vocabulary Graph Convolutional Network (VGCN). Additionally, we explore an ensemble using the aforementioned models. Our best model reaches an F1-score of 71.4% in Sub-task A and 67.3% for Sub-task B positioning our team in the upper third of the leaderboard. We release the code and experiments for our models on GitHub[1].

1 Introduction

The web and social network platforms, in particular, have become a significant part of our modern social lives. Sharing information, opinions, news, and jokes through these platforms with friends and family are now daily routines. One of the most prevalent forms of jokes in social networks are memes. Internet memes are small cultural units that are transformed, mixed, and shared using online platforms, often spreading in a viral manner (Milner, 2013). A prolific part are image-based memes, often available as templates that are accompanied by humorous or witty text. Unfortunately, a considerable proportion of the memes shared by Internet users are offensive or even hateful messages[2].

Detecting hate and offensive speech is a significant task for any online platform. Not only companies have this legal obligation in most countries, but also such language establishes a toxic environment that is detrimental to any online community on the long run. Hate speech can take multiple forms, but it is most frequently encountered as a disparaging message on the basis of a characteristic as race, gender, religion, and other criteria; Misogyny is one such frequent form specific to meme culture (Drakett et al., 2018; Phillips, 2012). Detecting hate speech is often a hard problem, even in an unimodal setting since the message often relies on the context, addresses current events, and incorporates cultural knowledge that cannot be easily incorporated into an automated model. The multi-modality of Internet memes increases the difficulty of the task since many memes can have a seemingly benign text that, contextualized with the associated image, becomes offensive or hateful.

Multi-modal hate speech detection had less attention in the research literature than traditional text-only methods. In the past two years, several datasets and challenges have addressed this by proposing detection tasks on meme-based data (Kiela et al., 2020; Gasparini et al., 2021; Miliani et al., 2020). Misogyny detection, as a subgroup of hate speech detection tasks, has also been more frequently encountered in research in recent years. The series of Automatic Misogyny Identification tasks proposed at IberEval 2018, EVALITA 2018, EVALITA 2020, and TRAC-2020 (Fersini et al., 2018a,b, 2020; Kumar et al., 2020; Bhattacharya et al., 2020) focused on the classification of tweets in English, Spanish, Italian, Bangla, and Hindi languages. In these tasks, researchers identified misogynous tweets and classified them as aggressive/non-aggressive or active/passive.

[1] https://github.com/readerbench/semeval-2022-task-5

[2] https://www.hmc.org.uk/blog/third-teenage-boys-admit-sending-receiving-racist-homophobic-content-online/

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 618 - 625
July 14-15, 2022 ©2022 Association for Computational Linguistics

Semeval-2022 Task 5: MAMI - Multimedia Automatic Misogyny Identification (Fersini et al., 2022) is a multi-modal classification task, aiming to detect misogynous memes by leveraging both image and text information. The task includes two sub-tasks: a binary identification of misogynous/non-misogynous memes (Sub-task A), and a multi-label classification distinguishing between the types of misogyny, namely: stereotype, shaming, objectification, and violence (Sub-task B).

In this paper, we present our contribution to this task by proposing two architectures based on the pre-trained multimodal UNITER (UNiversal Image-TExt Representation) model (Chen et al., 2020), as well as an ensemble from these two models. UNITER is an early fusion model pre-trained on large text-image datasets. UNITER leverages visual and location features extracted with Faster R-CNN (Ren et al., 2016), together with WordPiece encodings (Wu et al., 2016) derived from text tokens using a Transformer-based model (Vaswani et al., 2017). UNITER learns a generalizing representation for the text-image context by bringing the visual and text representations in a common embedding space. We use UNITER at the core of our two architectures: the first one enhances the pre-trained model with image sentiment features using a VGG-19 model (Vadicamo et al., 2017), while the second leverages graph convolutions on a co-occurrence graph (Kipf and Welling, 2016) built from an external dataset.

2 Background

Multi-modal tasks were traditionally associated with visual question answering (Goyal et al., 2017), image captioning (Gurari et al., 2020), audio-visual speech recognition (Paraskevopoulos et al., 2020), or cross-modal retrieval (Wang et al., 2016). With success of competitions like the Hateful Memes Challenge (Kiela et al., 2020), more research focused on multi-modal offensive classification. Pre-trained transformer models such as ViLBERT (Lu et al., 2019), VisualBERT (Li et al., 2019), LXMERT (Tan and Bansal, 2019), Oscar (Li et al., 2020), and others, dominated the competition leaderboard, either as stand-alone models or in large ensembles. While considering UNITER, Lippe et al. (2020) used an ensemble that placed them in the top 5 teams.

3 Method

We explored two ways of enhancing UNITER, first by adding a unimodal late fusion with a visual sentiment classifier, and second by using a multimodal early fusion with a modified Vocabulary Graph Convolutional Network (VGCN) (Lu et al., 2020; Paraschiv et al., 2021).

Dataset analysis and preprocessing

The training dataset contains 10,000 records, half of them misogynous. One misogynous record can have one or more of the four labels. The class distribution among the types of misogyny is as follows: shaming - 1,274, stereotype - 2,810, objectification - 2,208 and violence - 953 samples.

The text modality from the dataset was obtained, as far as we can tell, through OCR without any manual cleaning. This lead to the inclusion of date, times, mobile carrier names, Facebook user names or words on some unrelated objects in the image like *"Verizon LTE 4:41 PM Bikram Dec 11 at 12:31 AM"*. Additionally, many memes contain the watermark of the publishing website: "imgflip.com", "makeameme.org", "memez.com". Since most memes use full uppercase fonts, the letter casing was not reliable throughout the dataset and we choose to lowercase all training entries.

In our experiments, we tried two types of data cleaning techniques: one where we remove all timestamps and date mentions using the SUTime library (Chang and Manning, 2012) and a second with the supplementary step of removing all website mentions and Twitter usernames from the text.

Visual Sentiment-enhanced UNITER

Offensive texts, hate speech, and misogynous language are often correlated with negative sentiments, whereas the tone, context, and content is often highly loaded with polarized language (Ali et al., 2021; Gitari et al., 2015); as such, our intuition to enhance the UNITER model with a sentiment classifier. We focused only on image modality since large language models often already capture features required for text sentiment analysis. All images were classified using a pre-trained VGG-19 model (Simonyan and Zisserman, 2014) fine-tuned on the T4SA dataset (Vadicamo et al., 2017). The resulting 4,096 sized feature vectors were fused by concatenation with the UNITER's pooled output and classified through a fully connected layer in

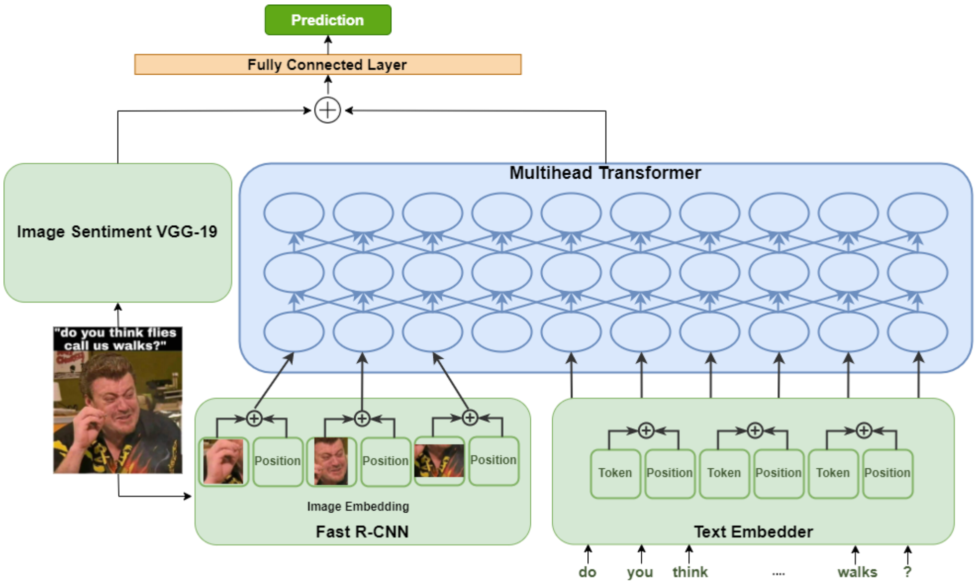

Figure 1: Illustration of the proposed UNITER-Sentiment model.

the classes for each sub-task (see Figure 1).

VGCN-enhanced UNITER

Graph enhanced BERT models have proven to be powerful for text classification (Mamani-Condori and Ochoa-Luna, 2021), even in multimodal tasks (Vlad et al., 2020). Using a GCN on a vocabulary graph (Kipf and Welling, 2016), the model can train an embedding layer to be fused with the Transformer embedding, before being processed by the multi-head attention layers. For our architecture, we employ a novel approach, namely to create a heterogeneous graph with nodes that represent text tokens and objects detected by the R-CNN layer, in the corresponding image.

A Kaggle dataset[3] containing 3,000 meme templates and their various possible text captions, totaling 533,827 text records, was used to build the aforementioned graph. The same R-CNN layer and object-token encoding as the UNITER model were considered to create a co-occurrence graph, having nodes as BERT-token-IDs and detected object-IDs, while edge values were computed using Point-wise Mutual Information (PMI). In contrast to (Lu et al., 2020), the obtained graph is independent of the training dataset and can be used for several tasks in the same domain.

In the training step, the UNITER image and text embeddings are fed through a GCN layer on top of the pre-built graph, thus generating a new embedding vector. The concatenated text, image, and graph embeddings are then processed by multi-head attention layers, while the pooled outputs are classified using a fully connected layer (see Figure 2).

Ensemble model

In order to leverage the learning of both proposed models, we also utilize an ensemble formed from multiple versions of both models, trained on different train/dev/validation splits, that are then combined via a soft voting scheme. The best performing trained versions evaluated on our test set were picked in the ensemble. For our final submission, we used the votes from 2 UNITER-Sentiment with UNITER-base, 2 with UNITER-large, and 7 UNITER-VGCN with UNITER-large. We chose these components of the ensemble based on the results for Sub-task B on our validation set, as seen in Table 1. Details on general hyper-parameters across all models are presented in the subsequent section. For UNITER-Sentiment models, we used 150 warm-up steps with a weight decay of 0.01, in contrast to only 120 warmup steps with a weight decay of 0.1 for UNITER-VGCN models. All UNITER-VGCN models in our ensemble had a graph embedding size of 16 and the edges had a minimum normalized PMI (Bouma, 2009) of 0.3. Since we trained each of the 11 ensemble components on a different train/dev/validation partition with the same seed, each configuration converged on different weights, with different performances. The results of all 11 models were combined through

[3] https://www.kaggle.com/zacchaeus/meme-project-raw

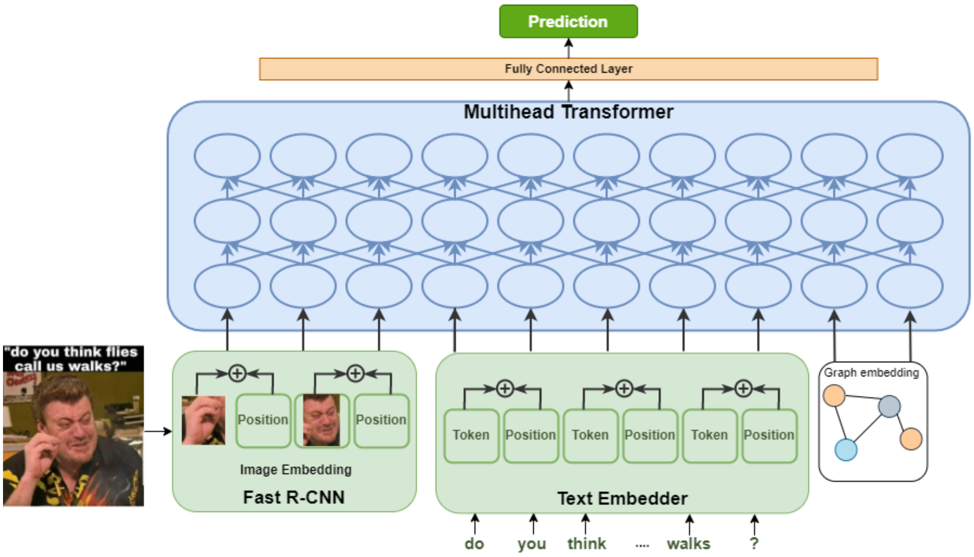

Figure 2: Illustration of the proposed UNITER-VGCN model.

a soft-voting scheme based on the average predicted probability for each label. Additional configurations were trained, but overall this ensemble had the best test performance among our submissions to SemEval-2022 Task 5.

3.1 Experimental Setup

Since the regions of interest (ROI) detection in the R-CNN layer[4] returns the probability for each rectangle, in our experiments, we use the minimum threshold for a ROI to be included in the dataset as an input hyperparameter. We experimented with both versions of UNITER: *UNITER-base* with 12 attention heads and 768 output dimensions, and *UNITER-large* with 24 attention heads and 1024 dimensions[5]. In order to mitigate class imbalance from Sub-task B, we used a weighted binary cross entropy loss using the class distribution in the training set. Other hyperparameter values were determined through grid search. Thus, the minimum confidence level for ROIs was set at 0.7, the learning rate at 1e^{-4} using an AdamW optimizer (Loshchilov and Hutter, 2017), and the maximum text length at 64 tokens. We experiment with GCN embedding sizes between 8 and 32, and selected 16 as the optimum value.

In order to further address the class imbalance between the misogyny sub-types, we experimented with oversampling from the minority classes, as an alternative to the loss re-weighting technique. Also, from our experiments, we learned that using an additional "non-misogynous" class besides the four misogyny types improved the model performance.

During training, we optimize the weighted-average F1-score (i.e., the F1-scores computed for each label, afterwards weighted by the label support) using early-stopping with a patience of 2 epochs.

4 Results

While considering the two proposed data preprocessing techniques, removing the websites from the textual information proved to decrease the performance of the models. Even though removing the source from the texts would provide a better generalization in a production setting, this information turned out to be a clue on misogyny content in this dataset.

Compensating for the unbalanced classes with oversampling from the minority classes proved to be less impactful than weighting the loss of the positive classes. Also, we noticed a strong tendency to overfit during our experiments. A common used mitigation technique is to augment the training data with small variations - e.g., data augmentation using similar words in the embedding space (Wang and Yang, 2015), as well as simpler replacement, swap, and deletion methods (Wei and Zou, 2019). None of the applied augmentation methods improved performance. We thus hypothesize that although these augmentations create similar meanings, innuendos get lost. For example, the phrase *"Her: Excuse me, I'm trying to put a **load** in the*

[4] https://github.com/MILVLG/bottom-up-attention.pytorch

[5] https://github.com/ChenRocks/UNITER

		Sub-			Sub-	
		task A			task B	
Model	**Precision**	**Recall**	**Weighted-F1**	**Precision**	**Recall**	**Weighted-F1**
UNITER-base+Sentiment$_1$	81.60%	63.95%	67.17%	68.41%	42.13%	61.34%
UNITER-base+Sentiment$_2$	83.80%	62.82%	66.16%	63.59%	44.72%	63.16%
UNITER-large+Sentiment$_1$	83.00%	63.17%	66.47%	57.67%	49.68%	64.69%
UNITER-large+Sentiment$_2$	**86.80**%	62.54%	66.13%	70.51%	45.00%	64.18%
UNITER-large+VGCN$_1$	78.80%	65.78%	68.59%	63.39%	48.87%	**65.89**%
UNITER-large+VGCN$_2$	78.40%	64.05%	66.78%	58.38%	47.06%	64.50%
UNITER-large+VGCN$_3$	78.40%	**68.89**%	**71.36%**	66.30%	46.54%	64.84%
UNITER-large+VGCN$_4$	77.60%	65.10%	67.70%	48.24%	**51.84**%	64.25%
UNITER-large+VGCN$_5$	86.60%	63.03%	66.74%	**70.91**%	42.76%	61.66%
UNITER-large+VGCN$_6$	79.20%	68.39%	71.12%	58.48%	50.62%	65.35%
UNITER-large+VGCN$_7$	81.60%	65.18%	68.50%	53.66%	50.18%	65.56%

Table 1: Results on our validation set for the Ensemble components.

Model	**Precision**	**Recall**	**Weighted-F1**
UNITER-base+Sentiment$_1$	72.60%	67.60%	68.86%
UNITER-base+VGCN	**86.20**%	61.66%	64.91%
UNITER-large+Sentiment$_1$	83.00%	63.16%	66.47%
UNITER-large+VGCN$_3$	78.40%	**68.89**%	**71.36**%
Ensemble	83.20%	66.88%	70.56%

Table 2: Results on the official test set for Sub-task A.

Model	**Precision**	**Recall**	**Weighted-F1**
UNITER-base+Sentiment$_2$	59.97%	46.43%	63.99%
UNITER-base+VGCN	63.99%	45.61%	63.39%
UNITER-large+Sentiment$_1$	57.67%	49.68%	64.68%
UNITER-large+VGCN$_1$	**66.30**%	46.54%	64.84%
Ensemble	63.19%	**50.65**%	**67.31%**

Table 3: Results on the official test set for Sub-task B.

Rank	**Team**	**Sub-task A Score**	**Team**	**Sub-task B Score**
1	SRC-B	83.4%	SRC-B	73.1%
2	DD-TIG	79.4%	TIB-VA	73.1%
3	beantown	77.8%	PAFC	73.1%
	UPB (our)	71.4%	UPB (our)	67.3%
	Baseline	65.0%	Baseline	62.1%

Table 4: Comparison for Sub-tasks A and B between the top 3 team results, our scores, and the competition baseline.

dishwasher Him: Same @gogomeme" would get an augmented counterpart *"Her: Excuse me, I'm trying to put a* ***burdened*** *in the dishwasher Him: Same @gogomeme"*. Changing the word "load", which in the context has a double meaning, loses the implicit misogynistic comparison of the woman to a dishwasher. Similarly, a text like *"my horny girlfriend on her period me"* augmented as *"my horny girlfriend on her deadline me"* would not improve the training.

Tables 2 and 3 are the best performances on the official competition test set for the models we trained. All modes were trained using the hyperparameters specified in the previous section.

Even though the performance on the binary classification from Sub-task A was comparable with

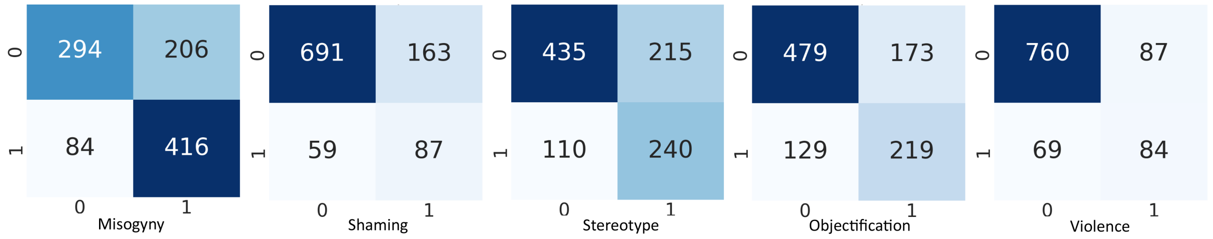

Figure 3: Per class confusion matrices for the ensemble model.

the best single model (UNITER-VGCN), it was slightly lower. As seen in Figure 3, our best system - the ensemble model - has the tendency to over-predict misogyny. Some erroneous predictions were driven by their aggressive language, for instance *"IF YOU'RE DATING MY DAUGHTER AND YOUR STUPID ENOUGH TO DO THIS I'M GOING TO KILL YOU!" depicting also that a domestic abuse victim was understandably detected as misogynous type "violence"*. Records 15566 and 15311 depict a face close up of a famous woman and were incorrectly labeled as misogynous, even if the text is replaced with a neutral or even positive one like "woman" or "best". This can be explained by the off-balance in the training data where the visual object "woman" is detected 2,262 times in the misogynous entries, and only 639 times in non-misogynous memes; similarly, "eyebrows" are four times more likely to appear in the misogynous class. However, memes like *"When you know it's a trap but you can't wait to take the bait"* that depict a woman body are not detected as misogyny since these memes require additional background knowledge in order to understand the real intent.

5 Conclusion

In this paper, we describe the architectures used in our submission at Semeval-2022, Task 5: MAMI - Multimedia Automatic Misogyny Identification. Our proposed models took the pre-trained UNITER-base and UNITER-large models and enhanced them with image sentiment features or, by using a GCN, with additional domain information from an external dataset. Our best model achieved an F1-score of 71.4 for Sub-task A and 67.3 in Sub-task B, seen in Table 4 in comparison to the top three results on each sub-task, arguing that these models can perform reasonable well in a multi-modal setting and that the generalization power of the UNITER pre-trained model was enhanced by integrating image object nodes in the co-occurrence graph. Also, we showed that our ensemble smoothed out the uneven performance caused by different train/dev data splits and improved the overall performance.

In terms of future work, we plan to continue our research into the automatic detection of abusive and hateful online content, and extend the experiments onto other pre-trained multi-modal models, as well as to attempt to improve their performance through task-adaptive pre-training (Gururangan et al., 2020). Even though Internet memes provide an interesting combination of textual message and image, they represent only one medium that can spread toxic messages. Studying other modalities like video or audio would help widen the understanding on how to detect and limit the spread of these undesired messages.

References

Muhammad Z. Ali, Ehsan-Ul-Haq, Sahar Rauf, Kashif Javed, and Sarmad Hussain. 2021. Improving hate speech detection of urdu tweets using sentiment analysis. *IEEE Access*, 9:84296–84305.

Shiladitya Bhattacharya, Siddharth Singh, Ritesh Kumar, Akanksha Bansal, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, and Atul Kr. Ojha. 2020. Developing a multilingual annotated corpus of misogyny and aggression. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 158–168, Marseille, France. European Language Resources Association (ELRA).

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30:31–40.

Angel X. Chang and Christopher Manning. 2012. SUTime: A library for recognizing and normalizing time expressions. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3735–3740, Istanbul, Turkey. European Language Resources Association (ELRA).

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. *arXiv:1909.11740 [cs]*. ArXiv: 1909.11740.

Jessica Drakett, Bridgette Rickett, Katy Day, and Kate Milnes. 2018. Old jokes, new media – online sexism and constructions of gender in internet memes. *Feminism & Psychology*, 28(1):109–127.

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.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018a. Overview of the evalita 2018 task on automatic misogyny identification (ami). *EVALITA Evaluation of NLP and Speech Tools for Italian*, 12:59.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2020. Ami@ evalita2020: Automatic misogyny identification. In *EVALITA*.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018b. Overview of the task on automatic misogyny identification at ibereval 2018. *Ibereval@ sepln*, 2150:214–228.

Francesca Gasparini, Giulia Rizzi, Aurora Saibene, and Elisabetta Fersini. 2021. Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content. *arXiv:2106.08409 [cs]*. ArXiv: 2106.08409.

Njagi Dennis Gitari, Zuping Zhang, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. Captioning images taken by people who are blind. In *European Conference on Computer Vision*, pages 417–434. Springer.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

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.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2020. Evaluating aggression identification in social media. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 1–5, Marseille, France. European Language Resources Association (ELRA).

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, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. *arXiv:2004.06165 [cs]*. ArXiv: 2004.06165.

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:2012.12871 [cs]*. ArXiv: 2012.12871.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

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.

Zhibin Lu, Pan Du, and Jian-Yun Nie. 2020. Vgcn-bert: Augmenting bert with graph embedding for text classification. *arXiv:2004.05707 [cs, stat]*. ArXiv: 2004.05707.

Errol Mamani-Condori and José Ochoa-Luna. 2021. Aggressive language detection using vgcn-bert for spanish texts. In *Brazilian Conference on Intelligent Systems*, pages 359–373. Springer.

Martina Miliani, Giulia Giorgi, Ilir Rama, Guido Anselmi, and Gianluca E. Lebani. 2020. *DANKMEMES @ EVALITA 2020: The Memeing of Life: Memes, Multimodality and Politics*, page 275–283. Accademia University Press.

Ryan M. Milner. 2013. Fcj-156 hacking the social: Internet memes, identity antagonism, and the logic of lulz. *The Fibreculture Journal*, pages 61–91.

Andrei Paraschiv, George-Eduard Zaharia, Dumitru-Clementin Cercel, and Mihai Dascalu. 2021. Graph convolutional networks applied to fakenews: Corona virus and 5g conspiracy. *University Politehnica of Bucharest Scientific Bulletin Series C-Electrical Engineering and Computer Science*, pages 71–82.

Georgios Paraskevopoulos, Srinivas Parthasarathy, Aparna Khare, and Shiva Sundaram. 2020. Multiresolution and multimodal speech recognition with transformers. *arXiv preprint arXiv:2004.14840*.

Whitney Phillips. 2012. This is why we can't have nice things: The origins, evolution and cultural embeddedness of online trolling. Accepted: 2012-12-07T23:12:32Z.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv:1506.01497 [cs]*. ArXiv: 1506.01497.

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

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

Lucia Vadicamo, Fabio Carrara, Andrea Cimino, Stefano Cresci, Felice Dell'Orletta, Fabrizio Falchi, and Maurizio Tesconi. 2017. Cross-media learning for image sentiment analysis in the wild. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, page 308–317. IEEE.

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.

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.

Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5005–5013.

William Yang Wang and Diyi Yang. 2015. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2557–2563, Lisbon, Portugal. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.