UIC-NLP at SemEval-2022 Task 5: Exploring Contrastive Learning for Multimodal Detection of Misogynistic Memes

Charic Farinango Cuervo and **Natalie Parde**
Natural Language Processing Laboratory
Department of Computer Science
University of Illinois at Chicago
{cfarin2, parde}@uic.edu

Abstract

Misogynistic memes are rampant on social media, and often convey their messages using multimodal signals (e.g., images paired with derogatory text or captions). However, to date very few multimodal systems have been leveraged for the detection of misogynistic memes. Recently, researchers have turned to contrastive learning solutions for a variety of problems. Most notably, OpenAI's CLIP model has served as an innovative solution for a variety of multimodal tasks. In this work, we experiment with contrastive learning to address the detection of misogynistic memes within the context of SemEval-2022 Task 5. Although our model does not achieve top results, these experiments provide important exploratory findings for this task. We conduct a detailed error analysis, revealing promising clues and offering a foundation for follow-up work.

1 Introduction

Hateful expressions on the Internet are widespread and mostly based on religion, gender, race, or physical attributes (Lippe et al., 2020). Such language exacerbates damaging societal problems such as racism, sexism, and other types of discrimination. In particular, misogynistic abuse has become very prevalent and poses a serious problem in cyberspace (Citron, 2014). Although extensive research on hate speech and misogyny has been conducted (Kumar et al., 2020), it has thus far centered on the analysis of text or images alone.

Memes, or social media images that communicate messages through the creative use of imagery understood to carry specific rhetorical value among members of a community, are a common platform for misogyny and hateful expressions. Approximately 78% of women use image-based social media multiple times per day (compared to 65% of men) (Fersini et al., 2022), making exposure to this harmful content alarmingly common. Classifying memes is challenging because of their multimodal interplay between image and text, as well as their region-specific interpretation, and existing multimodal approaches do not perform very well in the classification of hateful memes (Kiela et al., 2020). The underlying goal of SemEval-2022 Task 5 was to address this research gap, tackling the challenge of identifying misogynistic memes using multimodal data by inviting researchers to experiment with a variety of approaches.

Our team, *UIC-NLP*, investigated an adapted version of the Contrastive Language-Image Pre-training (CLIP) technique recently established and applied with success to numerous other tasks (Radford et al., 2021; Conde and Turgutlu, 2021; Galatolo et al., 2021). Our model, recorded on the leaderboard in the 6th leaderboard cluster under the OpenReview username "Charicfc," ranked 71st out of 83 participants and obtained an average F_1 score of 0.62. We analyze our model's predictions and offer insights and recommendations for improving upon it in the future.

2 Background

2.1 SemEval-2022 Task 5 (MAMI)

SemEval-2022 Task 5 was created to address the rise in the use of memes as a form of hate against women, which contributes to sexual stereotyping and gender inequality. The data used for this challenge is comprised of English-language memes collected from the web and manually annotated via crowdsourcing platforms. Each data sample has an image, its raw text in English (*transcript*), a binary annotation indicating the presence of misogyny, and (if applicable) the type of misogyny (shaming, objectification, stereotype, or violence). Our model addresses *Subtask 1*, which focuses on binary

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 785 - 792

classification of memes as misogynist or not misogynist. The output for a given sample is a confidence score for the predicted class. Although no approaches have sought to perform multimodal detection of misogynistic memes to date, we review work on classifying misogynistic expressions in text and multimodal classification of hateful memes in the following sections.

2.2 Identifying Misogyny in Text

Previous approaches in the related IberEval 2018 Automatic Misogyny Identification (AMI) task for misogyny detection in tweets (Fersini et al., 2018) leveraged statistical classification models, including variations of Support Vector Machines (SVM) (Nina-Alcocer, 2018; Pamungkas et al., 2018). Other recent work towards identifying misogyny in text has leveraged CNNs, LSTMs, and combined representations from models like BERT (Basile et al., 2019; Parikh et al., 2021). Recently, the Evalita 2020 AMI challenge best results were obtained by ensembles of fine-tuned BERT models (Lees et al., 2020).

2.3 Multimodal Classification of Hateful Memes

Multimodal learning has recently gained attention due to the poor performance of existing (unimodal) models on multimodal tasks (Lippe et al., 2020), with most recent solutions (context aware) employing neural architectures such as CNNs, RNNs, and Transformer-based attention models like BERT (Afridi et al., 2020; Modi and Parde, 2019; Parde, 2020). Although existing work on hate speech detection has largely relied on text-based features, this has gradually started to shift with the introduction of multimodal datasets (Lippe et al., 2020). In general, the focus has been shifted to BERT-based models (Afridi et al., 2020). In Facebook AI's Hateful Memes Challenge (Kiela et al., 2021) the top two models involved an ensemble of four Transformer-based models (Zhu, 2020; Muennighoff. 2020). The third and fourth place used fine-tuned VisualBERT (Velioglu and Rose, 2020) and UNITERT (Hossain et al., 2021) models. Aggarwal et al. (2021) extracted image features with a pretrained ResNet-152 model while passing the text data through Facebook's FastText encoder (Bojanowski et al., 2017), concatenating both feature vectors and passing them to a fully connected layer for classification. Due to its novelty, few solutions to date have explored the use of CLIP in challenging scenarios such as misogynistic memes classification. For this work, we trained a version of CLIP resembling Shariatnia's (2021).

3 System Overview

Our system architecture is a variation of OpenAI's CLIP model (Radford et al., 2021). We refer readers to the original paper for a detailed overview and illustrations of the architecture, and summarize it here. Inspired by the idea of usability and generality, CLIP uses a contrastive learning objective to build a joint visual-linguistic space for learning visual concepts from natural language supervision. We describe the various components of our architecture below.

3.1 Encoders

ResNet-50: OpenAI's best CLIP performance was achieved using a pretrained encoder which the authors called ViT-L/14@336px. In our case, we experimented with several pretrained image encoders. Although Shariatnia (2021) used ResNet-50 by default, we also considered ViT-B/16@224px, ViT-L/16@224px, and ViT-L/16@384px (Dosovitskiy et al, 2020). Nonetheless, we empirically determined that ResNet-50 (He, 2016), a deep CNN trained on more than a million images from the ImageNet database with an objective of classifying images into 1000 categories, yielded the best performance.

DistilBERT: OpenAI's CLIP used a modified Transformer to encode text. We instead used a lighter version of BERT (Devlin et al., 2018) called DistilBERT (Sanh et al, 2019) which Shariatnia (2021) also uses. BERT is a large Transformer-based language model that has achieved strong performance in many NLP tasks, and DistilBERT uses a process known as knowledge distillation to reproduce its behavior by training a smaller model to replicate its probability distributions across class predictions.

3.2 Learning Objective

Radford et al. (2021) introduced contrastive objectives as a mechanism for learning multimodal representations from raw images and paired descriptions. In essence, the contrastive objective seeks to learn a multimodal embedding space where image embeddings and text embeddings are

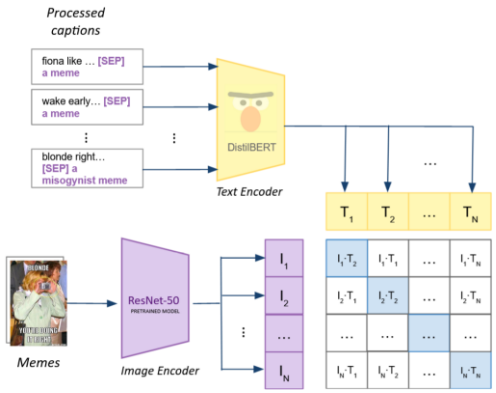

Figure 1. Learning stage. Adapted from Radford et al., 2021.

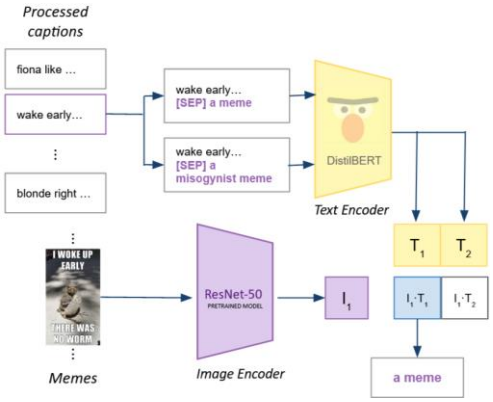

Figure 2. Evaluation stage. Adapted from Radford et al., 2021.

mapped to the same point if they describe the same thing, and different points otherwise. Cosine similarity is used to measure the distance between embeddings. Figure 1 shows the "logits" matrix obtained after applying the dot product between images and text embeddings. Each cell in the matrix (*logits*) is a measure of similarity between an image and a text caption in the dataset (N^2 pairs). It is expected that cells along the diagonal (N pairs), which contain the similarity between an image and its actual text, are maximized. Simultaneously, the $N^2 - N$ incorrect pairs should be minimized.

The *targets* for the images and texts where the similarities are maximum are obtained by computing dot products between embedding matrices. These are averaged and passed through a SoftMax, with the result being a *target* matrix where the diagonal is close to 1.0 and the other pairs are close to 0.0. The loss for images and texts is obtained by calculating the cross-entropy between the target and the logits matrix.

3.3 Training Procedures

The target output for the original CLIP model was the correct caption for an image. For example, one could input an image of a cat along with the captions: {"Image depicting a cat," "Image depicting a dog"} with the expectation that it would return the correct caption. We refer readers to the original paper for further implementation details.

We slightly modified this approach in our own model, such that the training text instead was the language content from the meme followed by the correct label. We separated the text and label using the [SEP] token as defined for BERT's next-sentence prediction objective. Therefore, we concatenated each instance with the following sentences depending on the example's class:

1. For class MISOGYNY: <text_1> + "[SEP] a misogynist meme"
2. For class NOT MISOGYNY: <text_1> + "[SEP] a meme"

When evaluating new samples using this architecture, each instance must be a paired image and text caption. Thus, we created two versions of each text caption: one for class MISOGYNY, and one for class NOT MISOGYNY. Figure 2 shows the procedure. The predicted label for the test instance was the one for which the model made its prediction with higher confidence.

4 Experimental Setup

4.1 Exploratory Analysis

Prior to conducting our experiments, we performed preliminary descriptive analyses on the training data, comprising 10,000 memes (image + paired text content) with a total vocabulary of 12,611 tokens. From these, 6,649 tokens only appear once. Figure 3 shows a word cloud for the 100 most frequent words in the corpus. Interestingly, we determined that the vocabulary for non-misogynist memes is much richer, including 9,285 compared to 7,348 unique words. We also observed that while the word "woman" is repeated 1,416 times in the misogyny class, it is repeated only 528 times in the not misogyny class.

Aiming to find greater insights we computed the Pointwise Mutual Information (PMI) score for all words given each class. PMI is a feature scoring metric that can be used to estimate the association between a feature and a class (it has also traditionally been used to identify collocations in text). A close association indicates which features

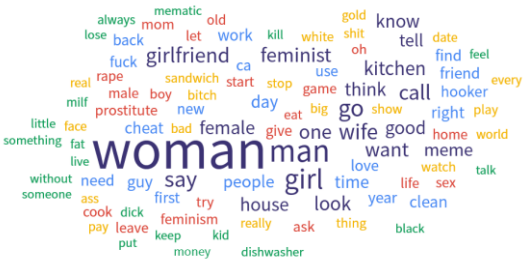

Figure 3: Word cloud of the 100 most frequent words in the corpus. *Warning: This image includes language that may be offensive or upsetting.*

(words) are more important for a class. PMI is computed using the following formula:

$$PMI(w,c) = \log\frac{p(w,c)}{p(w).p(c)} = \log\frac{p(w|c)}{p(w)} = \log\frac{p(c|w)}{p(c)} \quad (1)$$

Table 1 shows the 15 words with the highest PMI for each class. As shown, despite its frequency, the word "woman" carries less value when measured via PMI. Some swear words are classified as important for the MISOGYNY class, but not the NOT MISOGYNY class.

4.2 Preprocessing

A large challenge in this task was that since the text content from memes was extracted with an OCR tool, it contained substantial noise. Therefore, we created the following preprocessing pipeline that we applied to all texts before training and evaluating:

1. **Case Normalization:** All text was normalized by converting it to lowercase.
2. **Part of Speech (POS) Tagging:** POS tags were assigned to each word.
3. **Proper Name Removal:** Regular expressions were applied to convert some wordforms (e.g., urls) to generic tokens.
4. **Special Token Categorization:** Words belonging to several word sets of interest, including *celebrity names* and *profanity terms*, were kept.
5. **Lemmatization:** Words were converted to their base forms.
6. **Stopword Removal:** Highly frequent words (e.g., "the") were removed.
7. **Special Character Removal:** Non-alphabetical characters (e.g., digits or punctuation) were removed.

Aspects of this preprocessing pipeline were similar to those reported by Cardoza (2022) and Kovács et al. (2020). POS tagging allowed us to target the "proper noun" and "other" tags. By excluding these tags, we resolved many lingering issues following application of regular expressions, such as remaining usernames and gibberish words. Removing specific instances of these terms aided the model in avoiding overfitting to superfluous names or unknown tokens that were irrelevant to the overarching task of recognizing misogyny. Since certain terms removed by our POS filtering may carry importance to the task (e.g., certain celebrity names or swear terms), we also searched for these terms in several predetermined word sets and kept them if and when they were found.

Misogynist			Non-Misogynist	
Token	**PMI**		**Token**	**PMI**
dishwasher	0.660357		gold	0.591098
sandwich	0.599215		house	0.587787
rape	0.567609		cheat	0.470893
feminist	0.560566		clean	0.457903
fat	0.479573		call	0.456758
feminism	0.478848		people	0.430466
girl	0.453501		cook	0.407265
bitch	0.425832		game	0.40027
woman	0.403349		kid	0.394994
always	0.291434		girlfriend	0.391106

Table 1: Top 15 PMI scores. *Warning: This table includes language that may be offensive or upsetting.*

4.3 Experimental Settings

The training set provided by the task organizers was divided into training (80%) and validation (20%) sets of 8000 and 2000 examples respectively. We primarily adapted our hyperparameter settings from Shariatnia (2021), using AdamW (Loshchilov and Hutter, 2017) as our optimizer with an initial learning rate (LR) of 1e-3 and a scheduler to reduce the LR on plateau. The batch size is left at 32 and the maximum number of epochs was set to 4. The model converges quickly, so this early stop helps to control overfitting. The learning rates for the image and text encoders were left at 1e-4 and 1e-5 respectively. All of the texts were tokenized using the DistilBERT base model with a max number of tokens set to 200. We experimented with CLIP's temperature hyperparameter, finding that the best results were achieved with a value of 1e-0.2.

We selected the best epoch and the best hyper-parameters as measured by F_1 score and accuracy. In the test set evaluation, Subtask 1 systems were evaluated using macro-averaged F_1; thus, the final score is the mean of the F_1 for the two classes.

Development			
	Precision	**Recall**	**F_1**
Non-Misogynist	0.66	0.71	0.68
Misogynist	0.69	0.64	0.66
Accuracy			0.67
Evaluation			
	Precision	**Recall**	**F_1**
Non-Misogynist	0.73	0.43	0.54
Misogynist	0.60	0.84	0.70
Accuracy			0.64

Table 2: Results on the development (from training data) and evaluation (from test data).

5 Results

During development we used the training data provided by the MAMI task. We divided it into training (90%) and test (10%) sets, and measured performance using F_1, accuracy, precision, and recall on the 10% test set. We provide our results on the development data in Table 2. Once the official test set was released (*Evaluation* in Table 2), we computed the same metrics on that set. The macro-averaged F_1 as returned by the task organizers was 0.62, with the system ranking 71st in performance.

5.1 Quantitative Analysis

We briefly analyze our best models' results on the test set for Subtask 1. In particular, we observe a fairly low recall for the NOT MISOGYNY class (0.43), indicating that the system may be struggling to capture all members of this likely more diverse class. As further highlighted in the confusion matrix in Figure 4, the model classifies 28.6% of the NOT MISOGYNIST memes as MISOGYNIST. Thus, most errors were false positives.

5.2 Error Analysis

When manually analyzing the misclassified instances, we observed that they were diverse, containing cartoons, animals, people, drawings, and more. Some images contained text that was unrelated to the caption, creating additional noise. Examples of these images are shown in Figure 5. Several recurring themes occurred among false positives and false negatives, which we summarize below:

False Positives: Most images that were incorrectly classified as misogynistic were primarily dominated by one or more people. For images where humans were not present, the text often contained swear words or synonyms for "woman," some of which were offensive although not employed with purposeful misogyny.

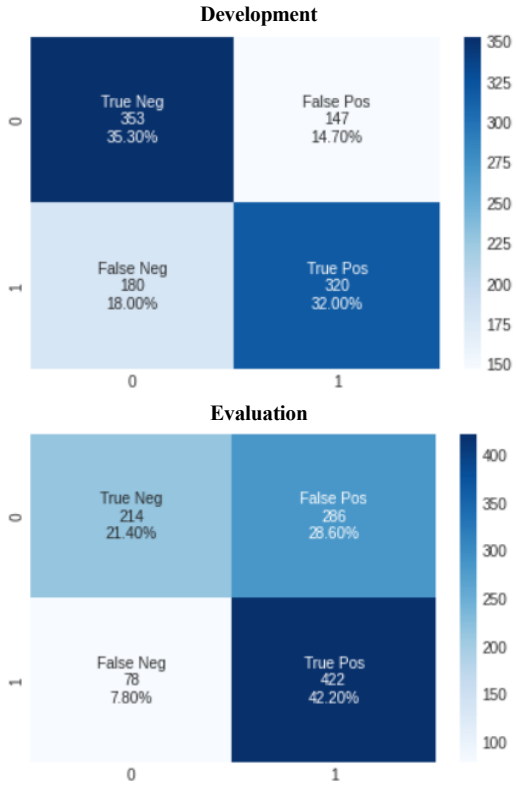

Figure 4: Confusion matrices for the development and evaluation data.

False Negatives: Most images that were incorrectly classified as not misogynistic contained cartoons, animals, storyboards or non-explicitly sexist images, although women were occasionally present.

To address these errors, we recommend actions leveraged in prior work for other tasks to improve the performance of future systems. For instance, Lippe et al. (2020) upsampled their dataset as a solution for poor performance on text confounders. Although the dataset they used (Facebook's Hateful Memes) was specially designed to introduce benign confounders, it might also work in this problem. Nozza et al., (2019) discuss biases introduced in the model by a set of identity terms that are frequently associated with the misogynistic class (e.g., "woman"). The authors propose to upsample the dataset with examples that have the identity terms for the alternate class.

6 Conclusion and Future Work

In this paper, we describe our system implementation for SemEval-2022 Task 5. Our

Figure 5: False positives (top 10) and false negatives (bottom 10). *Warning: This image includes content that may be offensive or upsetting.*

model ranked 71st out of 83 participants' teams on Subtask 1. We comprehensively investigate the use of a state-of-the-art multimodal contrastive learning approach for the classification of misogynistic memes. More experiments and tests should be done to improve the model's performance on this task. In particular, upsampling the dataset and addressing the possible biases caused by identity terms should be investigated. Finally, at an architectural level, our current system encodes images as one form of input and encodes paired text content and labels as another form of input, similarly to text encoding strategies used for unimodal sequence prediction tasks. Exploring joint encodings of image and paired text content as a single form of input, with only labels as the other form of input, may be an additional design avenue worth pursuing. Overall, it is our hope that this work motivates additional interest in contrastive learning solutions for multimodal misogynistic meme detection. We make our source code [1] available to other researchers to facilitate follow-up work by others.

7 Acknowledgements

We thank the anonymous reviewers for their comments and suggestions.

References

Tariq Habib Afridi, Aftab Alam, Muhammad Numan Khan, Jawad Khan and Young-Koo Lee. (2020, October). A multimodal memes classification: A survey and open research issues. In *The Proceedings of the Third International Conference on Smart City Applications* (pp. 1451-1466). Springer, Cham. https://doi.org/10.1007/978-3-030-66840-2_109

Apeksha Aggarwal, Vibhav Sharma, Anshul Trivedi, Mayank Yadav, Chirag Agrawal, Dilbag Singh, Vipul Mishra and Hassène Gritli. (2021). Two-way feature extraction using sequential and multimodal approach for hateful meme classification. *Complexity*, *2021*. https://doi.org/10.1155/2021/5510253

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana, Patti, Francisco Manuel Rangel Pardo, Paolo Ross and Manuela Sanguinetti. (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *13th International Workshop on Semantic Evaluation* (pp. 54-63). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/S19-2007

Piotr Bojanowski, Edouard Grave, Armand Joulin and Tomas Mikolov. (2017). Enriching word vectors with subword information. Transactions of the association for computational linguistics, 5, 135-146.

Maria Alejandra Cardoza Ceron. (2022). Using Word Embeddings to Analyze Protests News. arXiv preprint arXiv:2203.05875.

Danielle Keats Citron. (2014). Hate crimes in cyberspace. Harvard University Press.

Marcos V. Conde and Kerem Turgutlu. (2021). CLIP-Art: contrastive pre-training for fine-grained art classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3956-3960).

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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit and Neil Houlsby. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, Jeffrey Sorensen. (2022). SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification. In *Proceedings of the 16th International Workshop*

[1] https://github.com/charicf/MAMI-CLIP

on Semantic Evaluation (SemEval-2022). Association for Computational Linguistics.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. (2018). Overview of the Task on Automatic Misogyny Identification at IberEval 2018. *Ibereval@ sepln*, *2150*, 214-228.

Federico A. Galatolo, Mario G.C.A. Cimino and Gigliola Vaglini. (2021). Generating images from caption and vice versa via CLIP-Guided Generative Latent Space Search. arXiv preprint arXiv:2102.01645.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 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.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Casey A. Fitzpatrick, Peter Bull, Greg Lipstein, Tony Nelli, Ron Zhu, Niklas Muennighoff, Riza Velioglu, Jewgeni Rose, Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, Helen Yannakoudakis, Vlad Sandulescu, Umut Ozertem, Patrick Pantel, Lucia Specia, Devi Parikh. (2021, August). The hateful memes challenge: competition report. In *NeurIPS 2020 Competition and Demonstration Track* (pp. 344-360). PMLR.

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.

Ádám Kovács, Judit Ács, Andras Kornai, and Gábor Recski. 2020. Better Together: Modern Methods Plus Traditional Thinking in NP Alignment. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3635–3639, Marseille, France. European Language Resources Association.

Ritesh Kumar, Atul Kr. Ojha, Bornini Lahiri, Marcos Zampieri, Shervin Malmasi, Vanessa Murdock and Daniel Kadar. (2020, May). Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*.

Alyssa Lees, Jeffrey Sorensen and Ian Kivlichan. (2020). Jigsaw@ AMI and HaSpeeDe2: Fine-Tuning a Pre-Trained Comment-Domain BERT Model. In *EVALITA*.

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*.

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

Yatri Modi and Natalie Parde (2019). The Steep Road to Happily Ever after: an Analysis of Current Visual Storytelling Models. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, pages 47–57, Minneapolis, Minnesota. Association for Computational Linguistics.

Niklas Muennighoff. (2020). Vilio: state-of-the-art Visio-Linguistic models applied to hateful memes. *arXiv preprint arXiv:2012.07788*.

Victor Nina-Alcocer. (2018, September). AMI at IberEval2018 Automatic Misogyny Identification in Spanish and English Tweets. In *Ibereval@ sepln* (pp. 274-279).

Debora Nozza, Claudia Volpetti, Elisabetta Fersini. (2019, October). Unintended bias in misogyny detection. In *Ieee/wic/acm international conference on web intelligence* (pp. 149-155). https://doi.org/10.1145/3350546.3352512

Endang Wahyu Pamungkas, Alessandra Teresa Cignarella, Valerio Basile, and Viviana Patti. (2018). 14-ExLab@ UniTo for AMI at IberEval2018: Exploiting lexical knowledge for detecting misogyny in English and Spanish tweets. In *3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages, IberEval 2018* (Vol. 2150, pp. 234-241). CEUR-WS.

Natalie Parde (2020). And, Action! Towards Leveraging Multimodal Patterns for Storytelling and Content Analysis. In *Proceedings of the 2nd International Workshop on AI for Smart TV Content Production, Access and Delivery (AI4TV '20)*. Association for Computing Machinery, New York, NY, USA, 3. DOI:https://doi.org/10.1145/3422839.3423060

Pulkit Parikh, Harika Abburi, Niyati Chhaya, Manish Gupta and Vasudeva Varma. (2021). Categorizing Sexism and Misogyny through Neural Approaches. *ACM Transactions on the Web (TWEB)*, *15*(4), 1-31. https://doi.org/10.1145/3457189

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger and Ilya Sutskever. (2021,

July). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* (pp. 8748-8763). PMLR.

Victor Sanh, Lysandre Debut, Julien Chaumond and Thomas Wolf. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Moein Shariatnia. (2021). Moein-Shariatnia/OpenAI-clip: Simple implementation of openai clip model in PYTORCH. GitHub. Retrieved November 29, 2021, from https://github.com/moein-shariatnia/OpenAI-CLIP.

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

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