ASRtrans at SemEval-2022 Task 5: Transformer-based Models for Meme Classification

Ailneni Rakshitha Rao[1], Arjun Rao[2]
[1]Indian Institute of Technology, Gandhinagar
[2]Chaitanya Bharathi Institute of Technology, Hyderabad
rao_ailneni@alumni.iitgn.ac.in
ugs18088_ece.arjun@cbit.ac.in

Abstract

Women are frequently targeted online with hate speech and misogyny using tweets, memes, and other forms of communication. This paper describes our system for Task 5 of SemEval-2022: Multimedia Automatic Misogyny Identification (MAMI). We participated in both the sub-tasks, where we used transformer-based architecture to combine features of images and text. We explore models with multi-modal pre-training (VisualBERT) and text-based pre-training (MMBT) while drawing comparative results. We also show how additional training with task-related external data can improve the model performance. We achieved sizable improvements over baseline models and the official evaluation ranked our system 3^{rd} out of 83 teams on the binary classification task (Sub-task A) with an F1 score of 0.761, and 7^{th} out of 48 teams on the multi-label classification task (Sub-task B) with an F1 score of 0.705.

1 Introduction

Despite its unique advantages, social media is considered to be one of the harmful elements of society, if not monitored properly. It has become a medium to express hatred towards particular groups, especially women. Women have a strong presence online and have become victims of systematic inequality and discrimination which is reflected from the behavior offline. Violence has increased to the point where for many girls, abuse is a day-to-day reality. A landmark survey conducted by Plan International in more than 20 countries has revealed shocking accounts of escalating online violence against girls and women, with respondents exposed to explicit messages, pornographic photos, cyberstalking, and other forms of internet abuse. The most common type of online harm includes using abusive and insulting language, followed by deliberate embarrassment, body shaming, and threats of sexual violence.

One of the most popular communication tools in social media is a *meme*. It is a combination of image and text, created typically for humor. SemEval 2022 Task 5 is the first misogynistic meme detection challenge that incorporates several categories such as stereotyping, shaming, objectification and violence.

Pre-trained language models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), etc., have emerged as the state-of-the-art models for many NLP tasks such as text classification, machine translation, sequence tagging, etc., mainly due to their rich contextual embeddings. Hence, we chose a transformer-based architecture to fuse both visual and textual features. Two types of pre-training techniques are explored in this paper: text-based and multi-modal-based. Instead of directly using a pre-trained text transformer, we further train RoBERTa on task-related data and fine-tune the model with extended visual features extracted from a image classification network (e.g. ResNet). We also use ensemble learning to combine the results of the two pre-training techniques.

We achieved significant improvement over baselines in both the sub-tasks. We were ranked (1) 3^{rd} with an F1-macro score of 0.761 in sub-task A and (2) 7^{th} with a weighted F1 score of 0.705 in sub-task B among 83 teams. We release the code for models and experiments via GitHub [1]

The rest of the paper is organized as follows: Section 2 describes the challenge, followed by a brief literature survey. Section 3 explains the proposed approach in detail while section 4 presents the experimental details required to reproduce the results. Results and analysis are shown in section 5. Finally, conclusions are drawn in section 7.

[1] https://github.com/rak55/ASRtrans-semeval2022

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 597 - 604
July 14-15, 2022 ©2022 Association for Computational Linguistics

Label	No. of samples
Misogynous	5044
Shaming	1274
Objectification	2204
Violence	962
Stereotype	2844

Table 1: Distribution of training data.

2 Background

2.1 Problem Description

SemEval 2022 Task 5: MAMI: Multimedia Automatic Misogyny Identification (Fersini et al., 2022) consists of two sub-tasks: **Sub-task A** is a simple binary classification task to identify whether a meme is misogynous or not. **Sub-task B** is an advanced multi-label classification task, where memes are further classified among four categories namely *stereotype*, *shaming*, *objectification*, and *violence*.

2.2 Related Work

Multimodal data classification There are two types of approaches to multimodal data classification: late fusion and early fusion. Early works on multimodal data employed late fusion techniques such as combining major image features with bag-of-words-based text features (Tian et al., 2013). In this approach, two separate models are trained with images and text and their outputs are combined at a later stage. (Zhang and Pan, 2019) used a late fusion of CNN-based image features and RNN-based text features. The late fusion approach can be used even if one of the modalities is missing in the input. The disadvantage of this approach is, it fails to learn the interactions between different modalities. On the other hand, early fusion approaches use joint representations of images and text, thereby training a single model to learn within and across both modalities. With the development of pre-trained language models such as BERT, early fusion models like VisualBERT (Li et al., 2019), Vision and Language BERT (Lu et al., 2019), Visual-Linguistic BERT (Su et al., 2019), Multimodal Bitransformers (MMBT) (Kiela et al., 2019) and Learning Cross-Modality Encoder Representations from Transformers (Tan and Bansal, 2019) have quickly risen in popularity. However, early fusion models may perform poorly because using a single optimization strategy is sub-optimal for a model dealing with multiple modalities as explained in (Wang et al., 2019).

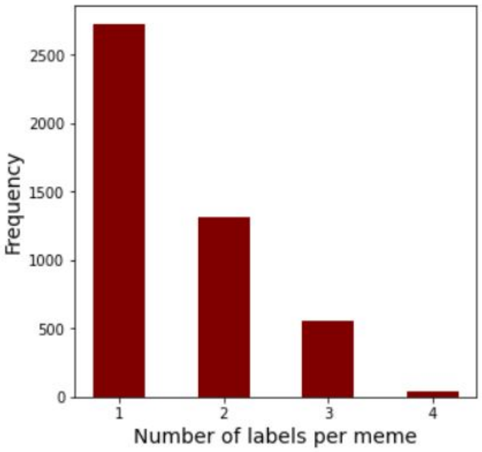


Figure 1: Frequency distribution of #labels per meme.

Multi-label classification There are mainly three different techniques to solve a multi-label classification problem (sub-task B) : Binary Relevance, Classifier Chains (Dembczyński et al., 2010) and Label Powerset method (Boutell et al., 2004). Binary Relevance treats each class independently and ignores label dependence. The Label Powerset method considers each combination of labels as a distinct class, thereby transforming a multi-label classification problem into a single-label problem. Classifier Chains connects binary classifiers in a chain such that the output of one classifier is treated as the input feature for the subsequent classifier.

3 System Overview

We explored models using both single-task and multi-task learning approaches. Since multi-task learning did not provide any significant improvement, our final model was trained using single-task learning. Consequently, we trained models using one modality (either text or image) and both modalities for the sake of comparison. We also compared the performance of models with and without multimodal pre-training.

3.1 Data

The training data provided for this task consists of 10,00 memes. The text transcription of memes along with the annotated labels for each image file is provided in a .csv file. Memes in the dataset are of random size. The distribution of different labels in the public training data is shown in Table 1. The frequency distribution of the number of labels per meme is shown in Figure 1. Each meme is labeled as either misogynous or non-misogynous

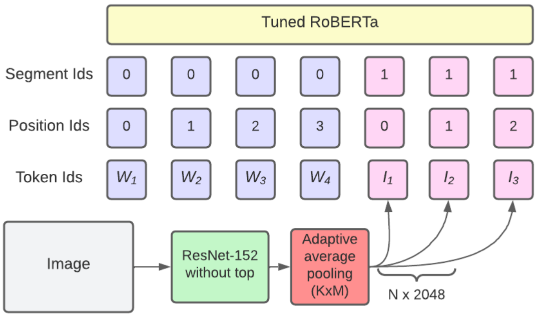


Figure 2: Architecture of the proposed model.

and in turn, misogynous memes are further classified among potential overlapping categories. 20% of data is randomly sampled from the training data for validation. The final test data provided for evaluation consists of 1000 memes.

3.2 Single modality approach

3.2.1 Text-based approach

We tested the following text-based models: i) A bidirectional LSTM model with Glove embeddings. ii) Pre-trained RoBERTa-base model fine-tuned with a classification head. iii) Ernie-2.0 (Sun et al., 2019), fine-tuned in the same way as RoBERTa-base model. RoBERTa model is further trained with external data and fine-tuned with training data for classification.

3.2.2 Leveraging External data

RoBERTa (Liu et al., 2019) has 12 layers, 12 heads, and a hidden layer dimension of 768. It is trained with masked language modeling on five datasets: BookCorpus, English Wikipedia, CC-News, OpenWebText, and Stories. We initiated it with pre-trained weights and trained on sexism-based datasets using the HuggingFace library. Sexism-based datasets are used since the language and content present is similar to our training dataset. (Grosz and CONDE-CESPEDES, 2020) introduced a labeled dataset consisting of sexist comments made in the workplace. We took 624 comments labeled 'sexist' out of 1137 comments present in this dataset. (Singh et al., 2021) presented a dataset to determine the use of sexism in English sitcoms from which, we extracted 1631 sexist text instances. From EXIST2021 dataset (Rodríguez-Sánchez et al., 2021-09), sexist text instances in English language are extracted. Altogether, we curated 5049 sexist text instances for training. Since the acquired dataset is small, we did not train the RoBERTa model from scratch. We call this model *tuned* RoBERTa (tRoBERTa) for the entirety of this paper.

3.2.3 Image-based approach

We fine-tuned several large-scale image feature extraction networks such as VGG-16 (Liu and Deng, 2015), ResNet-50, ResNet-152 (He et al., 2015) and Vision transformer (Dosovitskiy et al., 2021) for both the sub-tasks.

3.3 Multimodal input

We implemented the late fusion approach mentioned in section 3.2.1 by combining image features from VGG-16 and text features from a BiLSTM model. Besides, we also trained the state-of-the-art multimodal models such as VisualBERT and MMBT described in section 2.2.

VisualBERT is an integration of BERT and pre-trained object proposal system, Faster-RCNN. Similar to BERT, VisualBERT is pre-trained using two language model objectives: 1) Masked language modeling, where part of the input text is masked and predicted using contextual visual and text tokens. 2) Sentence-image prediction where the model determines if the text data matches the image or not. By combining image and text regions through a transformer, VisualBERT aims to learn useful alignments between them. For this purpose, it uses unordered visual embeddings extracted from an object detector, each corresponding to a bounded region in the image. In addition to BERT inputs, VisualBERT takes visual embeddings, visual token type ids, and visual attention masks as input. We fine-tuned VisualBERT for the task which is pre-trained on the VQA task with COCO dataset. Image features are extracted from a ResNeXt-based Faster RCNN, pre-trained on Visual Genome dataset.

MMBT combines representations from large language models and state-of-the-art convolutional neural networks in a straightforward way. It employs a BERT-base uncased model (12-layer 768-dim) trained on English Wikipedia. It uses N separate image embeddings extracted from the ResNet-152 network with average pooling over K x M grids (N = K x M). The dimension of each embedding is 2048. MMBT maps these image embeddings to BERT's token space using a set of randomly initialized mappings. The output of the [CLS] token in the last layer of BERT is given to a dense layer for classification. In our system, we

Model	Task A	Task B
BiLSTM + Glove emb.	0.585	0.571
RoBERTa (single-task)	0.646	0.629
RoBERTa (Multi-task)	0.648	0.631
Ernie-2.0	0.643	0.630
VGG-16	0.639	0.610
ResNet-50	0.617	0.608
ResNet-152	0.631	0.611
Vision Transformer	0.624	0.609
VGG-16 + BiLSTM	0.641	0.625
MMBT	0.725	0.669
VisualBERT	0.723	0.673
MMBT with tRoBERTa	**0.751**	**0.700**
Avg. Ensemble	0.761	0.705

Table 2: F1 score of all the major models on test dataset for both the sub-tasks. Avg. Ensemble is the weighted average of both the models.

use the tuned RoBERTa (tRoBERTa) described in section 3.2.2 instead of the BERT model used in the original implementation. The architecture of our proposed model is shown in Figure 2. This model is fine-tuned with a classification head on top by optimizing focal binary cross-entropy loss for multi-label classification.

3.4 Ensemble Learning

Multimodal models such as VisualBERT and MMBT differ in their training procedures and the datasets on which they are trained. Hence, they may focus on different aspects of the input. So, it is a good practice to combine the results of these two models to learn accurate representations. There are several ways to combine them: we can concatenate embeddings of different models and project them to a low dimensional space for prediction, but this will require high computational power. Instead, we can train these models independently and later combine their predictions. In a Weighted-Average Ensemble, results are obtained by taking a weighted average of the predictions. In this case, the weights are obtained by grid search on the validation dataset. Another way to combine the predictions is the Voting Ensemble method, where the class predicted by the majority of the models is considered as the final output. We experimented with both approaches and found that the weighted average method yields better results than the voting ensemble method.

Rank	Team	F1-macro
1	SRC-B	0.834
3	DD-TIG	0.794
5	NLPros	0.771
6	**ASRtrans**	0.761
61	Baseline_Text	0.640
62	Baseline_Image	0.639
79	Baseline_Image_Text	0.543

Table 3: Comparison of our sub-task A results with those on leaderboard.

4 Experimental Setup

We used pytorch (Paszke et al., 2019) and HuggingFace library (Wolf et al., 2019) for training and inference. All the models are trained on Google Colab. AdamW (Loshchilov and Hutter, 2019) optimizers with learning rates of 2e-5 and 5e-5 are used for training Visualbert and MMBT models respectively. Other text models such as RoBERTa and Ernie-2.0 also use AdamW optimizer with a learning rate of 2e-5. The maximum length of the text is limited to 50. We chose a batch size of 32 for training all the models. All the hyperparameters are tuned on the validation set which is 20% of the training data.

4.1 Data prepocessing

An input image is resized to 256 x 256 and then center-cropped to 224 x 224, followed by normalization before passing into MMBT. Text transcriptions of memes are cleaned to get rid of any URLs, HTML tags, and punctuation. Subsequently, they

Rank	Team	F1
1	SRC-B	0.731
7	NLPros	0.720
8	QMUL	0.713
14	**ASRtrans**	0.705
41	Baseline_Hierarchial	0.621
48	Baseline_Flat	0.421

Table 4: Comparison of our sub-task B results with those on leaderboard.

are annotated with the [CLS] token in the beginning before passing to the model. Removing stopwords showed a slight deterioration in the performance. Hence, we did not remove them.

4.2 Data Augmentation

Data augmentation is widely used to generate slightly variant larger datasets from the existing smaller ones. Since one recurring issue among all the models we trained is overfitting, three types of data augmentation techniques are tested to address this issue. We applied contextual augmentation for labeled sentences as proposed in (Kobayashi, 2018). In this method, we replace the words in a sentence with words predicted by a bi-directional language model. We also tested the back-translation approach proposed in (Sennrich et al., 2016) and easy data augmentation (EDA) techniques described in (Wei and Zou, 2019). We observed that while back-translation and contextual augmentation slightly improved the performance of the model, the use of EDA degraded it.

4.3 Loss functions

We trained our model for sub-task A with binary cross-entropy loss, whereas for sub-task B we used focal loss. Focal loss (Lin et al., 2020) is a form of cross-entropy loss, but it is dynamically scaled. It is computed as:

$$FL\left(p_t\right) \ = (1-p_t)^{\gamma} \log\left(p_t\right)$$

By setting $\gamma > 0$, we are reducing the relative loss for easy, well-classified examples ($p_t > 0.5$). For prediction, we used a threshold of 0.35 as it produced better results on the validation set.

5 Results and Analysis

We have tested transformer and non-transformer-based models with features extracted from text, images, and a combination of text and images as input. We experimented with models that employ text and multimodal pre-training like MMBT and VisualBERT respectively. The results of these experiments are summarized in Table 2. We can infer from the results that pre-training transformer-based language models give better results compared to LSTM-based text models. This performance can be attributed to their rich contextual embeddings.

System	F1
MMBT with RoBERTa	0.682
+ tuned RoBERTa	0.690
+ focal bce loss	0.696
+ threshold at 0.35	0.700

Table 6: **Sub-task B (dev)**: Incremental analysis of our system.

No. of img embeds	F1
1	0.751
3	0.743
4	0.739
5	0.740

Table 5: Comparison of results of sub-task A with different number of image embeddings.

Our model (MMBT with tuned RoBERTa) performs slightly better than VisualBERT without any multimodal pre-training. This is because VisualBERT is pre-trained on Microsoft's image annotation dataset COCO and most real-world multimodal inputs are not as straightforward as a caption describing an image. For the model to adapt to the target domain, we need additional pre-training with the data of interest. In our case, instead of extra multimodal pre-training, we just trained the text model (RoBERTa) for a few epochs using external data. This is computationally more efficient and flexible. Furthermore, this approach can be easily applied to inputs with missing modalities.

The performance of our model on the development set with a different number of image embeddings is shown in Table 5. Using a single image embedding gave better results than using multiple ones. Also, an incremental analysis of our system is shown in Table 6. This shows the importance of further training RoBERTa and focal loss. Besides, combining VisualBERT and our model in an ensemble gave an additional performance boost of around 1% (F1 score) in sub-task A and 0.5% in sub-task B.

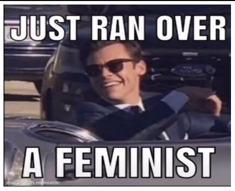
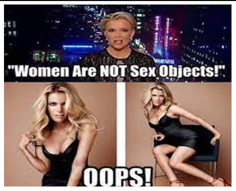
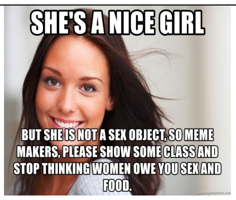
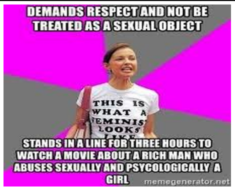
Meme			 SHE'S A NICE GIRL BUT SHE IS NOT A SEX OBJECT, SO MEME MAKERS, PLEASE SHOW SOME CLASS AND STOP THINKING WOMEN OWE YOU SEX AND FOOD.	 DEMANDS RESPECT AND NOT BE TREATED AS A SEXUAL OBJECT THIS IS WHAT A FEMINIST LOOKS LIKE STANDS IN A LINE FOR THREE HOURS TO WATCH A MOVIE ABOUT A RICH MAN WHO ABUSES SEXUALLY AND PSYCOLOGICALLY A GIRL memegenerator.net
Pred. label	stereotype, violence	stereotype, objectification	shaming, stereotype, objectification	objectification
Org. label	stereotype, violence	stereotype, objectification	non-misogynous	shaming, stereotype

Table 7: Sample test predictions of our model.

Comparison of the results of our model with the top submissions on the leaderboard for sub-task A and sub-task B are reported in Table 3 and 4 respectively. The official baselines for sub-task A are fine-tuned USE sentence embeddings (only text as input), fine-tuned VGG-16 model pre-trained on ImageNet dataset (only image as input), and concatenation of the above two models (image plus text as input). Baselines for sub-task B are: i) flat multi-label classification model with the concatenation of USE sentence embeddings and VGG-16 model features. ii) a hierarchical multi-label model based on USE sentence embeddings. Our model outperforms the best baseline model by 12% in sub-task A and 8.4% in sub-task B.

Error analysis of our model is shown in Table 7. The first two columns are examples of correctly classified memes whereas the last two columns are the incorrectly classified ones. The Model fails to classify the third meme as non-misogynous because it overlooks the word ’not’. This has been a common problem in language models such as BERT since they don’t understand negation well, resulting in incomplete syntactic knowledge. In the last meme, the usage of the words *sexual objects* most likely misled the model into classifying it as *objectification* without taking into account ’not’. As this is a common error, in future, special methods should be devised to help models overcome it.

6 Conclusion and Future Work

We conducted experiments with models trained on features extracted from text, images, and both modalities combined for meme classification. To evaluate the text-based approach, we trained language models such as RoBERTa, Ernie-2.0, etc., and RNN-based models. Subsequently, we concluded that the transformer-based models produced best results. We showed that further training the RoBERTa model (tuned RoBERTa) with task-related data improves the performance. This tuned RoBERTa model combined with the features from ResNet-152 without multimodal pretraining performs slightly better than VisualBERT. Furthermore, an ensemble of these two models gave an additional performance boost. In the future, we plan to test other problem transformation approaches such as Classifier Chains and their variants for multi-label classification in order to accurately model label dependence and hierarchy. Additionally, we plan to train VisualBERT from scratch on meme datasets to see if there is an improvement in its performance.

References

Matthew R. Boutell, Jiebo Luo, Xipeng Shen, and Christopher M. Brown. 2004. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771.

Krzysztof Dembczyński, Weiwei Cheng, and Eyke Hüllermeier. 2010. Bayes optimal multilabel classification via probabilistic classifier chains. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML’10, page 279–286, Madison, WI, USA. Omnipress.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*.

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.

Dylan Grosz and Patricia CONDE-CESPEDES. 2020. Automatic Detection of Sexist Statements Commonly Used at the Workplace. In *Pacific Asian Conference on Knowledge Discovery and Data Mining (PAKDD), Wokshop (Learning Data Representation for Clustering) LDRC*, Singapour, Singapore.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. *CoRR*, abs/1909.02950.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. *ArXiv*, abs/1805.06201.

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. *CoRR*, abs/1908.03557.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2020. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327.

Shuying Liu and Weihong Deng. 2015. Very deep convolutional neural network based image classification using small training sample size. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 730–734.

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.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *CoRR*, abs/1908.02265.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, Laura Plaza Morales, Julio Gonzalo Arroyo, Paolo Rosso, Miriam Comet, and Trinidad Donoso. 2021-09. Overview of exist 2021: sexism identification in social networks.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Smriti Singh, Tanvi Anand, Arijit Ghosh Chowdhury, and Zeerak Waseem. 2021. "hold on honey, men at work": A semi-supervised approach to detecting sexism in sitcoms. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 180–185, Online. Association for Computational Linguistics.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. VL-BERT: pretraining of generic visual-linguistic representations. *CoRR*, abs/1908.08530.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2019. Ernie 2.0: A continual pre-training framework for language understanding.

Hao Tan and Mohit Bansal. 2019. LXMERT: learning cross-modality encoder representations from transformers. *CoRR*, abs/1908.07490.

Lexiao Tian, Dequan Zheng, and Conghui Zhu. 2013. Image classification based on the combination of text features and visual features. *Int. J. Intell. Syst.*, 28(3):242–256.

Weiyao Wang, Du Tran, and Matt Feiszli. 2019. What makes training multi-modal networks hard? *CoRR*, abs/1905.12681.

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 6383–6389, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Han Zhang and Jennifer Pan. 2019. Casm: A deep-learning approach for identifying collective action events with text and image data from social media. *Sociological Methodology*, 49(1):1–57.