Project report for ANLP final project

Sebastian Wilharm, 773088
University of Potsdam
swilharm@uni-potsdam.de

Abstract

The task of this project is applying binary classification to the Multimedia Automatic Misogyny Identification dataset to detect misogyny in memes. The idea of our approach is classifying text and image separately and comparing their performances with those of a combined approach. Results show much higher performance for the image-only classifier, which is also not beaten by the multimodal approach.

1 Introduction

In modern times the internet has very high social significance and especially the younger generations spend a significant amount of time each day online, amongst others on social media platforms. As such, trends on these platforms can strongly influence the worldview of its consumers, so it is important to monitor their contents and to inhibit problematic content and behavior as early as possible. Due to the sheer amount of content being produced every minute however, it has become unfeasible for humans to moderate these platforms. We therefore need to rely on machine learning models to identify and flag offensive content.

On platforms like reddit, 9gag or 4chan a popular medium of content is the meme, an image with an overlaid caption. These are typically intended to convey jokes, many of which relatively harmless but some of them being of offensive nature. The goal of this research is training a model that can automatically detect misogyny in memes as a binary classification problem. Because memes are images with texts on them, this task allows the usage of both image classification and text classification techniques. Our idea was to unimodally classify only text or only image and to compare their results with a combined multimodal approach. Image classification turns out to perform better on this task than text classification and an approach combining their results was not able to outperform the unimodal image-only approach. An early multimodal approach, where image and text are directly fed into a shared model was also planned but did unfortunately not pan out.

2 The Task

2.1 Task Description

The task of this project is applying binary classification to the *Multimedia Automatic Misogyny Identification* (MAMI) dataset to detect misogyny in images with text on them, so called memes. Misogyny is defined as ’hatred of, aversion to, or prejudice against women’ (Misogyny, n.d.) and has been shown to be highly prevalent online, especially on websites where memes are being exchanged (Paciello et al., 2021). This content is however produced at such a rapid rate, that monitoring all new entries and identifying problematic imagery is not possible. It is therefore of interest to be able to automatically detect memes displaying this kind of hate speech to combat this rising issue.

2.2 The Dataset

The dataset consists of images, depicting memes, together with a csv file which, for each image, contains the classification labels and the text on the image. It is therefore not necessary to perform any character recognition. The data was collected from the web and annotated through the use of a crowdsourcing platform. The data is already split into a training set of 10,000 images and a test set of 1,000 images. In each set, exactly half of the data is labeled as misogynous and the other half as non-misogynous. Looking at the word frequencies after stop word removal, the words *girl*, *women* and *woman* are amongst the most commonly appearing ones, at a much higher rate than usual (Table 3; Google Ngram Viewer, n.d.), so clearly the images were selected with a specific bias.

Unfortunately, very little about the curation of the dataset was published by the task coordinators. The data was originally taken from the web, so no reliable conclusions can be drawn about the speaker specifications. We do not know age, gender or ethnicities and while all the data is in English, the speakers could be anywhere between a native speaker and someone who barely speaks the language. The number of speakers is presumably high, but can also not be guaranteed. A very similar issue applies to the annotators, since the data was annotated via a crowdsourcing platform. The curators did not publish any information about the demographic of the annotators. It is also impossible to check if the data was annotated by a multiple different people or if one single curator possibly annotated all the data which would introduce significant bias. Figure 1 shows an example of an image labeled as non-misogynous and an image labeled as misogynous.

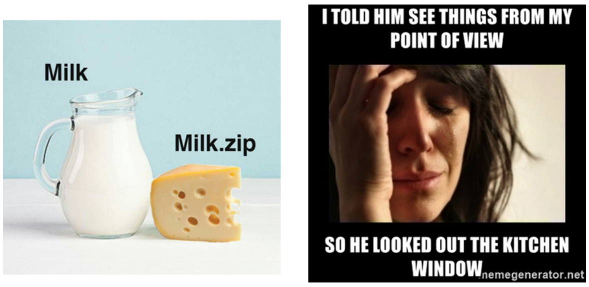
Milk
Milk.zip
I TOLD HIM SEE THINGS FROM MY POINT OF VIEW
SO HE LOOKED OUT THE KITCHEN WINDOW
memegenerator.net

Figure 1: Example of labeled images from the dataset Left: non-misogynous; Right: misogynous

2.3 Ethical Considerations

Automating this type of task is necessary as the pace of content creation is simply too high for human-led control mechanisms. Besides society as a whole, which would be helped in reducing misogyny, the main benefactor would be the platforms that themselves have an interest in policing hateful content, be that for moral reason or legal reasons as certain types of content moderation have become mandatory in areas like the EU.

While the data collected from the web guarantees anonymity and confidentiality of the speakers, as no identifying information is included, they had no opportunity to opt in or out of the study and did not give informed consent. As for potential for harm, many of the images in the dataset are highly offensive and might be triggering to some, especially when it comes to topics such as abuse and rape.

Since the curation of the dataset is not publicly documented, it puts to question its bias both in curation and in annotation. The images show an above average amount of women and the texts include words such as *girl*, *woman* and *women* at above average frequency. As we expected, one of the main discriminating factors the image model seems to be the presence of a woman in the picture, even though, obviously, pictures of women are not inherently more misogynous. It is therefore likely that the dataset is biased and a model trained on it could cause harm if it classifies all images of women as offensive, thus even increasing the effects of sexism. Imagine a world in which a woman uploads a picture of herself and gets flagged as posting misogynous content.

Considering a good number of the speakers likely intentionally created misogynous content, a model capable of detecting it could also be used to intentionally find, share and promote this type of content to gain notoriety in circles where this type of discrimination is appreciated.

3 Approach & Results

3.1 Planned Approach

We planned to both attempt unimodal approaches, where only either image or text are used as input for a model, as well as a multimodal approach, which combines the two. Fersini et al. (2019) and Kiela et al. (2020) show not only that unimodal approaches can be competitive with multimodal ones, but also compare so called early multimodal approaches, those that feed image and text representations into a shared classifier, with late multimodal approaches, which first separately classify image and text and then combine their respective outputs.

For text classification we settled on using BERT (Bidirectional Encoder Representations from Transformers), a commonly used language representation model (Das et al., 2020; Fersini et al., 2019; Kiela et al., 2020) that had been shown to discriminate data well for binary misogyny classification (Fersini et al., 2021).

For image classification we decided on using Resnet since it appeared to be used quite commonly with good results (Aggarwal et al., 2021; Kiela et al., 2020).

Besides the idea of combining classification outputs of the two unimodal approaches (late multimodal), there was also the plan to use a multimodal approach which directly combines both types of input data. A commonly mentioned pretrained model for this type of task was VisualBERT, which is used for joint representations of texts and images, trained on the COCO dataset (Das et al., 2020; Kiela et al., 2020; Muennighoff, 2020). We planned to also try this approach to compare to that of the late multimodal one.

3.2 BERT

Bidirectional encoder representations from transformers (BERT) is a pretrained model published in 2018 by Google with prevalent use for text classification tasks. There exist two versions of BERT, $\text{BERT}_{\text{BASE}}$, which consists of 12 transformer layers with hidden size 768 and 12 self-attention heads and $\text{BERT}_{\text{LARGE}}$, which consists of 24 transformer layers with hidden size 1024 and 16 self-attention heads. We decided that $\text{BERT}_{\text{BASE}}$ would be sufficient for our task. BERT is pretrained on the BooksCorpus with 800M words and English Wikipedia with 2,500M words. BERT uses WordPiece embeddings and expects as input a sequence of tokens, with two special tokens. [CLS] represents the first token of every sequence while [SEP] separates sentences from each other. The maximum sequence length BERT can accept is 512 tokens with short sequences being padded with another special token: [PAD]. BERT can typically be fine-tuned on task-specific data within a couple epochs. (Devlin et al., 2019)

3.3 Unimodal Text Classification

To use BERT in our task, we first imported the models from Hugging Face through the transformers library. The texts are tokenized with the BertTokenizer with a maximum sequence length of 512. The task-specific dataset is built and split into training and validation sets by a ratio of 80:20. The pretrained BertModel is then finetuned for 3 epochs on the training set and evaluated after each epoch on validation and test set. We used cross entropy loss as the loss function and the Adam optimizer with a learning rate of 1e-6. The accuracies reached after each epoch can be seen in Table 1. While training accuracy still improves in the third epoch, there is only little improvement in validation accuracy and any further training would likely just lead to overfitting.

Epoch	Train Acc.	Val. Acc.	Test Acc
1	0.599	0.720	0.528
2	0.776	0.775	0.535
3	0.848	0.780	0.552

Table 1: Training Accuracies for Text classification

3.4 Residual Neural Networks

A residual neural network is a deep neural network which uses the concept of "shortcut connections", i.e., skipping one or more layers. This combats two problems common in deep networks: vanishing gradients and the degradation problem, which leads to an increase in training error. The idea behind the skipping is to make every few layers fit to a residual function, instead of all layers together fitting the whole complex function. It is implemented by adding the output of one layer onto the input of another layer skipping multiple layers. The resulting networks are the so called Resnet models, where the number indicates the number of layers, e.g., Resnet-18 has 18 layers. Wide Resnets are a special kind which use a widening factor to allow for a shallower network. Wide Resnet-50-2 consists of 50 layers, but each twice as wide as in a regular Resnet. (He et al., 2016)

3.5 Unimodal Image Classification

We tried a large number of different image classification models, namely Resnet-18, AlexNet, VGG-16, Inception v3, ResNext5032x4d, Wide ResNet-50-2 and Regnet_x_400mf, but found Wide ResNet-50-2 to have the best results. We performed hyperparameter tuning and found best performance with stochastic gradient descent with a learning rate of 0.00001 and a batch size of 16. Image size does not seem to play a large role with sizes of 64x64, 128x128 and 256x256 all reaching very similar results. We settled on 256x256. We applied image transformations, such as rotation, center cropping and color jitter to avoid overfitting on the training data.

The model is finetuned for 5 epochs on the training set and evaluated after each epoch on the test set. The accuracies reached after each epoch can be seen in Table 2.

Epoch	Train Acc.	Test Acc.
1	0.562	0.560
2	0.616	0.585
3	0.647	0.586
4	0.659	0.604
5	0.667	0.611

Table 2: Training Accuracies for Image classification

Word	All	Mis.	Non-Mis.	Pred. Mis.	Pred. Non-Mis.
girl	0.6323	0.4689	0.8007	1.0693	0.3846
women	0.5653	0.762	0.3626	0.6786	0.5012
like	0.491	0.6301	0.3475	0.3907	0.5478
woman	0.4612	0.6594	0.2568	0.6169	0.373
make	0.3943	0.5129	0.2719	0.4935	0.338
meme	0.3794	0.381	0.3777	0.5346	0.2914
get	0.3422	0.337	0.3475	0.4113	0.303
girlfriend	0.3422	0.2198	0.4683	0.1851	0.4312
memegenerator.net	**0.3273**	**0.3077**	**0.3475**	**0.8842**	**0.0117**
imgflip.com	0.305	0.3077	0.3022	0.1234	0.4079
men	0.2976	0.3224	0.2719	0.2879	0.303
memes	0.2976	0.3077	0.2871	0.3496	0.2681
quickmeme.com	**0.2752**	**0.2931**	**0.2568**	**0.7608**	**0**
bitch	0.2678	0.4689	0.0604	0.3701	0.2098
got	0.2529	0.2052	0.3022	0.3085	0.2214

Table 3: Test set word frequencies in all texts, misogynous texts, non-misogynous texts, texts predicted to be misogynous and texts predicted to be non-misogynous

3.6 VisualBert

VisualBert consists of a stack of transformer layers that align input text and input image to model vision-and-language tasks. It integrates BERT to process the text inputs and object proposal systems such as Faster-RCNN to process the image inputs. It was mainly developed with tasks such as visual question answering and visual commonsense reasoning in mind, but has been shown to perform well on classification tasks as well (Das et al., 2020; Kiela et al., 2020; Muennighoff, 2020). VisualBert is pretrained on the COCO dataset and expects as inputs the tokenized text, once again preprocessed by the BertTokenizer and object proposal regions of a detector model. (Li et al., 2019)

3.7 Multimodal Classification

For a late multimodal approach, we combined the outputs of the two unimodal approaches, predicting as misogynous any meme that either the text classifier or the image classifier classified as misogynous. The resulting F1 score is 61.3%.

For an early multimodal approach, we had chosen to use VisualBert. While several papers talk about using VisualBert as a classification model, none of them go into detail how exactly this was done or have publicly available code. Documentation however suggested that the visual question answering model with a classification head on top would be the best candidate. We imported BertTokenizer and Faster-RCNN to preprocess the data exactly as done in the publicly available pretraining script and added a linear layer to the end to reshape the output from its previous 768 output dimensions to 2 dimensions to retrieve a binary classification output. When training however, the model showed no learning effect whatsoever. We tried with a number of different optimizers and loss functions to no avail. After double checking for programming errors, we concluded that we were likely using the model wrong or made a mistake with the preprocessing steps. Unfortunately, we could not solve this issue in time so we were essentially left with a coin flip classifier. We therefore decided not to include it in our implementation results but wanted to still describe the attempted approach.

4 Results & Evaluation

4.1 Unimodal Text Classification

Evaluation of the text classifier on the test set resulted in an accuracy of 54.8%, and a macro-averaged F1 score of 53.3%. Considering the dataset is evenly split between the two classes, this is fairly poor result.

It is difficult to determine why a neural network is making errors, as the decision making is obscured by a gigantic number of weights that influence its decision making. We generated frequency maps of word occurrence after lowercasing and stop word removal, which can be seen in Table 3. The table list the 15 most frequent words in all test texts and their respective frequencies in the labeled positive class and negative class and the predicted positive class and negative class. There are a few words whose frequencies do not match up between predictions and golden labels, such as *girl* and *like*, but of particular interest are the two lines printed in bold *memegenerator.net* and *quickmeme.com*, both websites to quickly generate images with overlaid text, which show no bias in the golden labels, but are significantly more common in texts predicted to be misogynous. The transformer model used in the project is of course not as simple as a bag of words model and this kind of analysis approach is therefore flawed but it gives some potential suggestions. Looking at the word frequencies in the training set, shows that these URLs are indeed much more common in positive training images than in negative training images, so this might be where the error originates from.

4.2 Unimodal Image Classification

Evaluation of the image classifier on the test set resulted in an accuracy of 66.3%, and a macro-averaged F1 score of also 66.3%. This is a much better result.

To understand what the model may be using for its decision making and where potential errors come in, we looked at images that were labeled non-offensive but predicted offensive as exemplary shown in figure 2 and vice versa as exemplary shown in figure 3. Noticeable is the much higher ratio of women in images that were predicted to be misogynous so this seems like a likely criterion that influenced the models choice. Memes that are only offensive through their text on the other hand, are generally classified as non-misogynous.

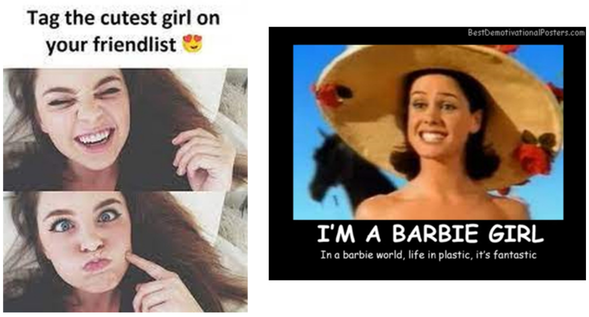
Tag the cutest girl on your friendlist 😍
BestDemotivationalPosters.com
I'M A BARBIE GIRL
In a barbie world, life in plastic, it's fantastic

Figure 2: Images labeled non-offensive but predicted offensive

4.3 Multimodal Classification

Evaluation of the late multimodal classifier on the test set resulted in an accuracy of 62.8% and a macro-averaged F1 score of 61.3%. We tried a number of different ways to combine the outputs of the unimodal classifiers, but none proved as good as the unimodal image classifier.

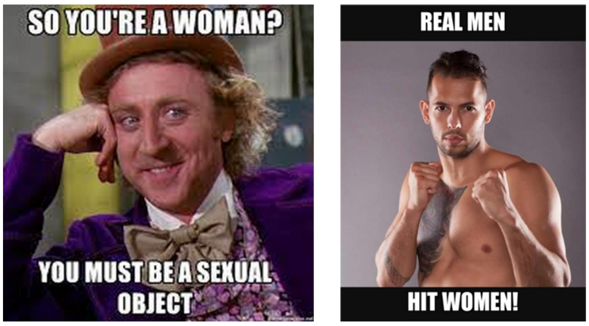

Figure 3: Images labeled offensive but predicted non-offensive

The performance of the combined classifier is displayed in table 4. The models agreed on 499 images, almost exactly half of the test set, and were correct in 70% of the agreement cases, performing especially well on non-misogynous images. They disagreed on the other 501 images, with the image model being correct 62.5% of the time. Interestingly the text model performed better on the non-misogynous disagreements, while the image model performed better on the misogynous disagreements.

Agreements	Non-Mis.	Mis.	
Both correct	215	135	350
Both wrong	62	87	149
	277	222	499
Disagreements	Non-Mis.	Mis.	
Text correct	125	63	188
Image correct	98	215	313
	223	278	501

Table 4: Agreements and disagreements of the models

5 Conclusion

The multimodality of the task allowed for a two-pronged approach, classifying texts and images separately. While BERT quickly converged on the training data, its performance on the test set leaves to be desired. The image classifier Wide Resnet-50-2 on the other hand performed much better and showed the best results out of the approaches. Combining their results did not yield a better result than image classification on its own. Unfortunately, the early multimodal approach of using VisualBert did not work out, as it would've provided another potentially strong model.

References

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.

Abhishek Das, Japsimar Singh Wahi, and Siyao Li. 2020. Detecting hate speech in multi-modal memes. *arXiv e-prints*, pages arXiv–2012.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. pages 4171–4186.

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

Elisabetta Fersini, Luca Rosato, Antonio Candelieri, Francesco Archetti, and Enza Messina. 2021. Deep learning representations in automatic misogyny identification: What do we gain and what do we miss? In *Eighth Italian Conference on Computational Linguistics, CLiC-it 2021, Milan, Italy, January 26-28, 2022*.

Google Ngram Viewer. girl,women,woman [online]. n.d. accessed: 2022-03-30.

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.

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.

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 e-prints*, pages arXiv–1908.

Misogyny. n.d. In *Merriam-Webster's collegiate dictionary*. Accessed: 2022-03-30. [link].

Niklas Muennighoff. 2020. Vilio: state-of-the-art visio-linguistic models applied to hateful memes. *arXiv e-prints*, pages arXiv–2012.

Marinella Paciello, Francesca D'Errico, Giorgia Saleri, and Ernestina Lamponi. 2021. Online sexist meme and its effects on moral and emotional processes in social media. *Computers in human behavior*, 116:106655.

A Learning Objectives & Personal Contribution

This project was my first project in the field of natural language processing and also my first project using neural networks. I therefore learned a lot through having to apply the techniques taught in class without a provided framework and instead doing every step on my own. I found it very interesting to use and finetune pretrained models, as they come with their own quirks and challenges. Image classification is the part of the project I enjoyed particularly as I had never done it before and it seemed more elusive to me than classifying texts. After initial shared planning, we decided to split the work into three parts, with Nellie focusing on image classification, Aleksandra focusing on text classification and me, Sebastian, focusing on the multimodal approach with VisualBert. As such this means that my work did unfortunately not make it into the final product, as, despite many attempts, the VisualBert model never worked out. I was able to bring my programming experience in again at the end when it came to preparing and restructuring all our work and results for the submission.

I think we overall worked well as a team, compensating for each others weaknesses, though I wish I would've carried more of the project when the current circumstances complicated the situation for my Russian teammates significantly. Luckily we managed to finish the project in time and it was an overall good experience.