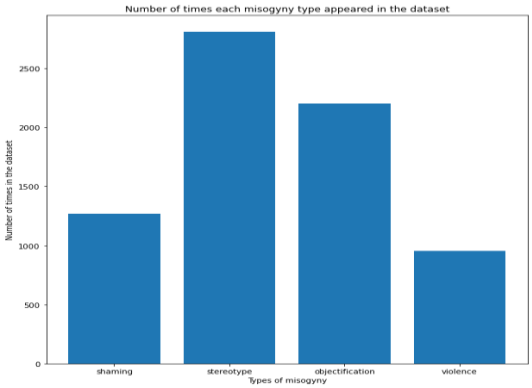
Multimedia Automatic Misogyny Identification (MAMI) Report

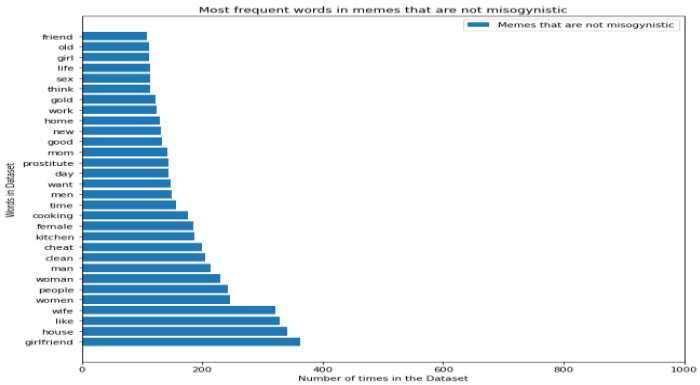
Disclaimer: Disturbing content included.

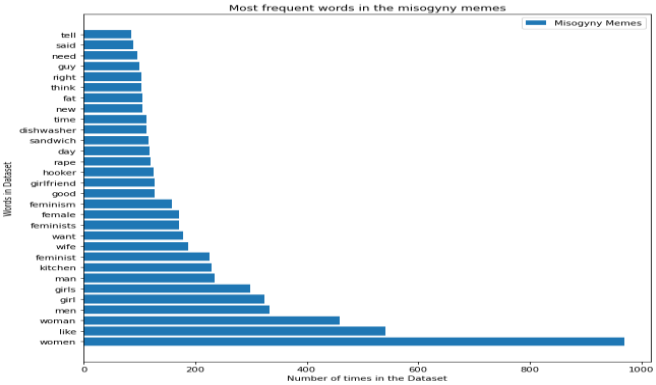
I. Exploratory Data Analysis



The MAMI dataset includes a label for misogyny valued as one if the meme is classified as misogynistic and zero if the meme is not misogynistic. Out of 10000 data, 5000 of them are labelled as misogynistic. Based only on these misogynistic data, there are furthermore four more labels (shaming, stereotype, objectification, violence) that characterize the type of misogyny. The first graph on the left depicts the number of times each misogyny type appeared in the dataset of 5000 misogynistic data. Turns out 1271 are shaming, 2810 are stereotyping, 2201 are objectifying and 953 violent. (may have more than one type)

Another useful information comes from the frequency of the words that are included in either category of misogynistic or not. The second graph on the left below depicts the most frequent words in memes that are not misogynistic, and the right graph depicts the same for the misogynistic memes. We conclude that the word 'women'/'woman' is the most frequent word and was used almost 1500 times in misogynistic memes. Also, other significant words in this category are 'men, girl, kitchen, feminist, wife, hooker, rape, dishwasher', which means that these words will have a strong affect in classification to misogyny category. The second graph shows that in non-misogynistic category the most frequent word is girlfriend, which appeared 360 times. There are many words that appear in both categories, but the frequency in non-misogynous category is lower. The paradox is that in this category there are words like 'cheat, kitchen, cooking, prostitute', that are expected to be used only in misogynistic memes.





II. Approach

First, I created the train and the test set using as dependent variable the misogyny label and as independent variables the pixels of the image. The test set was created with train_test_split and as test size I tested 2500, 2000 and 1000 data. The first model I created was based on the images of the memes. From tensorflow keras I loaded the images with 'rgb' colour and reduced the size to 90x90. Then I converted them in arrays and used each pixel of each image as input for the model. Lastly, several models were tested and the results suggested to use the logistic regression. (Some of the models could not run due to constraints of CPU and due to time constraints).

The second model was based on the text of each meme. The train and test sets were created the same way as it was described previously. For the text extraction, the TfidVectorizer was used. As parameters, the words were set to lowercase and stop words included all the English stop words and several words that were extracted manually. Lastly, several models were tested such as SVM model, Naïve Bayes, Decision Tree, KNN, logistic regression and neural network. The results suggested that the SVM model with C=1 and kernel 'rbf' and the logistic regression without intercept produced the best results.

III. Results

Confusion Matrix for model based on Images from Memes

Predicted:	**Non-Misogynistic**	**Misogynistic**
True Non-Misogynistic	301	208
True Misogynistic	206	285

Confusion Matrix for model based on Text from Memes

Predicted:	**Non-Misogynistic**	**Misogynistic**
True Non-Misogynistic	400	92
True Misogynistic	123	385

The confusion matrices above state that image-based model predicted correctly 301 and incorrectly 206 on the class non-misogynistic and predicted correctly 285 and incorrectly 208 on the class misogynistic. Similarly, the confusion matric on the right that was based on the text from the meme predicted 400 correctly and 123 incorrectly in class non-misogynistic and 385 correctly and 92 incorrectly in class misogynistic.

Classification Report for model based on Images from Memes

	Precision	***Recall***	***F1-score***
Non-misogynistic	0.59	0.59	0.59
Misogynistic	0.58	0.58	0.58
Accuracy			0.59
Macro Avg	0.59	0.59	0.59
Weighted Avg	0.59	0.59	0.59

Classification Report for model based on Text from Memes

	Precision	***Recall***	***F1-score***
Non-misogynistic	0.76	0.81	0.79
Misogynistic	0.81	0.76	0.78
Accuracy			0.79
Macro Avg	0.79	0.79	0.78
Weighted Avg	0.79	0.79	0.78

The classification reports are based on a logistic model (the SVM model provided same results). The Image-based model predicts a new meme accurately as misogynous or not with a percentage of 59%, while the text-based model almost with 80%. Precision is the ratio of correctly predicted observations of a class to the total predicted observations in that class and recall is the ratio of correctly predicted to the total actual observations in that class. The F1 score is the weighted average of precision and recall. Lastly, the Mean squared value for the test set is 0.50 for the first model and 0.2 for the second and for the train test 0.10 for both.

IV. Discussion

Based on the results, we reject the model based on the image of the meme, as it is based on patterns that pixels create and is unlikely to provide useful information. The prediction accuracy is low (59%) and the methods that were tested for combining the two modes did not provide any useful information. On the contrary, as it was expected by our initial hypothesis the text includes most of the information of the meme. The accuracy of the predictions (80%) is satisfactory and as it is expected the model could not get better without the appropriate combination of text and image. Another problem could arise from the selection of memes for the dataset, as they are selected from specific websites and the model could be applied only to memes from those websites. Lastly, based on Mean squared error there are not indications of overfitting.