Multimedia Automatic Misogyny Identification (MAMI) Report

Disclaimer: Disturbing content included.

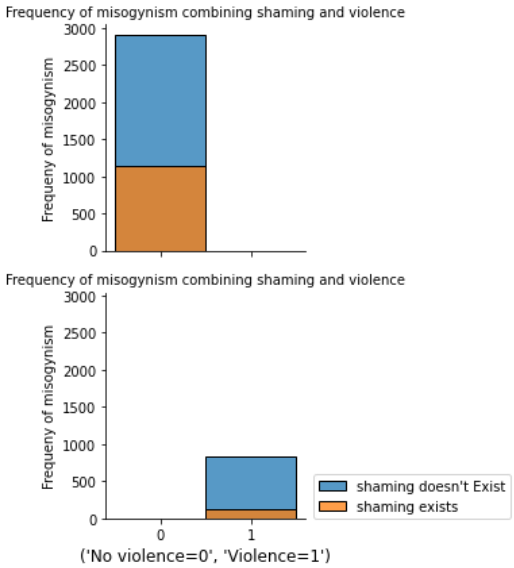
Name: Athanasios Memtsas

Student ID: f3352115

I. Exploratory Data Analysis

The main features of this assignment are from the text and the images. Although I firstly trained only the text to target misogyny,i still tried to add the information of the picture and compare the 2 confusion matrixes etc. But firstly lets explain the balance and plots :

Both classes have exactly 5000 (so no plot needed) ! But with this we cant say much,so lets take the words that show toxicity the most.So I tried to visualize it by combining shaming and violence only for misogynous meme's:



The output isn't quite as expected.It can be seen that for most misogynism cases only the 1/3 has violence and from those the most have not been annoted with shaming.

From wordcloud graphs lets observe the words with the biggest frequency in misogynous memes:

We see some words containing sites and none important details ,that will be extracted.

II. Approach

For text: I firstly cleaned every meme sentence from words over 20 characters,under 2 and removed punctuation,arithmetics,sites that memes are extracted from ,stopwords and applied stemming (that made my results worst so I removed it).

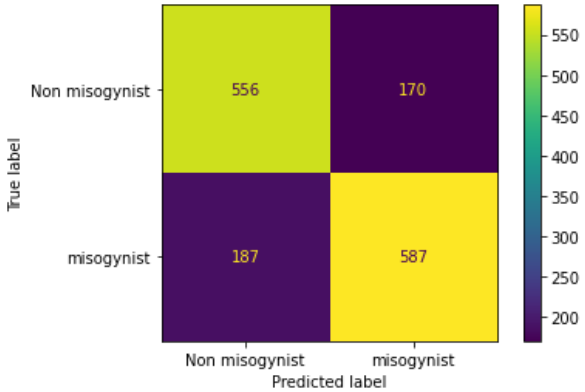
Next I vectorized each meme's text via tfidf split the data and trained with models from sklearn like LinearRegression that doesn't contain parameters,Random forest by increasing the number of trees in the forest and the max_depth of each tree to improve the model.I also tried a keras (simple neural network as presented in the slides) wih 100 epochs,10^(-7) learning rate,restoring best weights and batch size equal to 128.
Instead of Tfidf ,I also tried genism Word2vec whith size=300 window=3 and workers=10 but the similar function and other functions didn't seem to present the results I wanted ,so I preferred tfidf.

For the pictures: Loading and resizing each picture to (120x120) so it can be accepted as equal feature vectors for each picture and with image_to_array from keras the picture is represented by RGB . So splitting to Train pictures and Test, and reshaping from 4D to 2D we can now classify each meme via picture training to misogynous or not.

I applied the same models as above with close parameters and extreme but the accuracy seemed to be distributed among the same space.Also i tried a keras but my pc couldn't train 8.800 pictures so lightly so I stopped and dropped it.

III. Results

The best texting classifier was random forest:



Neural network had accuracy 0.68 and imbalanced f1 score and as we see random forest had better predictions with accuracy=0.78!!
For the images the best accuracy I could reach was 0.52 and imbalanced f1 score with random forest classifying.

IV. Discussion

In the end i couldn't classify memes from text+image in the same feature list and we prefer the text predictions as they are more accurate even if "an image is 1000 words" as they say .The only thought for the future I have,is maybe encode the image phases and spectrals with the vector of each memes test.Maybe apply nudity detection for each meme and compare misogynity frequency on nude pictures.