QiNiAn at SemEval-2022 Task 5: Multi-Modal Misogyny Detection and Classification

Qin Gu, Nino Meisinger, Anna-Katharina Dick
University of Tuebingen
{firstName.lastName}@student.uni-tuebingen.de

Abstract

In this paper, we describe our submission to the misogyny classification challenge at SemEval-2022. We propose two models for the two subtasks of the challenge: The first uses joint image and text classification to classify memes as either misogynistic or not. This model uses a majority voting ensemble structure built on traditional classifiers and additional image information such as age, gender and nudity estimations. The second model uses a RoBERTa classifier on the text transcriptions to additionally identify the type of problematic ideas the memes perpetuate. Our submissions perform above all organizer submitted baselines. For binary misogyny classification, our system achieved the fifth place on the leaderboard, with a macro F1-score of 0.665. For multi-label classification identifying the type of misogyny, our model achieved place 19 on the leaderboard, with a weighted F1-score of 0.637.

1 Introduction

Even though women are as much present online as men, some online spaces, such as microblogging websites, are still male dominated. Misogynistic jokes and memes are inevitably somewhat common in certain parts of the Internet and have the potential to perpetuate harmful ideas about gender or instill false ideas and expectations. Notably, the way they are spread is often through various modalities, most commonly visual and textual. Therefore, it would be useful to have a system that could automatically detect if certain combinations of texts and images are misogynistic or not. This is not a trivial task however, since misogyny can manifest in many different forms such as stereotyping, objectifying or threatening violence against women. A crucial difficulty in this multi-modal classification task is also the interplay between text and image. Some memes may appear harmless if only either the image or text are viewed separately. The Multimedia Automatic Misogyny Identification (MAMI) challenge at SemEval-2022 (Fersini et al., 2022) seeks to find solutions to solve this task.

We propose two models that automatically detect English misogynistic memes and classify the type of problematic ideas they perpetuate.[1] For simple binary misogyny detection, we created an ensemble model that makes predictions based on majority voting on two text-based and two image-based classifiers using partly hand-crafted features such as age, gender and nudity classification. For task B we relied only on text information and used a transformer based approach, creating a RoBERTa model (Liu et al., 2019) that classifies the type of misogyny that is perpetuated in the memes.

2 Background

Automatically identifying misogynistic texts has been explored in the past. In 2020, this task was proposed as an EVALITA shared task, using Italian tweets (Fersini et al., 2022). So far, the research in the area of misogyny detection has mostly focused on pure text data from social media, specifically Twitter (Anzovino et al., 2018; Frenda et al., 2019). In the current multi-modal task however, competitors were given the opportunity to explore classification using both visual and textual data given. Singh et al. (2020) used a multi-modal multi-task learning system with BERT (Devlin et al., 2018) features to extract textual information and ResNet features to handle image classification. A similar approach proved to be useful in another related shared task, in which Tamil memes should be identified as trolling or not (Suryawanshi and Chakravarthi, 2021).

[1] The code is made available at https://github.com/cicl-iscl/SemEval-2022_Multimedia_Automatic_Misogyny_Identification.

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 736 - 741
July 14-15, 2022 ©2022 Association for Computational Linguistics

shaming	THE FACE YOU MAKE WHEN TRUMP HAS THE CLASSY FOREIGN CHICK AND YOU'RE MARRIED TO KERCHAK imgrip.com
stereotype	creator made native women beautiful, to hide all that crazy
objectification	When my girlfriend is trying to have a serious conversation with me TITTY
violence	ROSES ARE RED, VIOLETS ARE BLUE IF YOU DON'T SAY YES, I'LL JUST RAPE YOU quickmeme.com
shaming + stereotype	CAN'T TELL IF THIS IS AN UGLY HIPPY CHICK OR REALLY PRETTY HIPPY BOY quickmeme.com
stereotype + objectification	Keeping your dishwasher clean will make it last longer Take care of your appliances
objectification + violence	inglip.com IS RAPING A PROSTITUTE A THEFT?

Table 1: Examples of text transcriptions that fit the four categories of misogyny and combinations of multiple labels.

The provided data set includes both images and text transcriptions of memes. The two subtasks we participated in were structured as follows: In subtask A, the memes should simply be identified as misogynous or not misogynous. Subtask B posed a more advanced challenge, as memes should additionally be classified as being part of four overlapping categories. These categories specify the type of misogyny expressed in the meme: stereotypes, shaming, objectification and violence. Shaming memes will insult women's appearance, stereotypes perpetuate harmful ideas about (groups of) women, objectification reduces women to their sexuality or body and violence downplays or advocates for violence against women. Examples of text transcriptions of these categories can be found in Table 1. The dataset for training includes 10000 memes with all of these labels, with exactly half of the memes being classified as misogynous and half as harmless. 2810 memes in the training set perpetuate stereotypes, 2202 objectification, 1274 shaming, and 953 violence. Many of the misogynistic memes have therefore overlapping categories, being classified as two or even three or all types of misogynous.

3 Subtask A: Binary misogyny classification

3.1 System Overview

For the binary classification of memes we experimented with a number of classification algorithms for both the image and the text transcriptions, as well as with a variety of different features. We decided to examine the benefits of an ensemble model that uses more traditional forms of text classification, as optimized ensemble models have been shown to perform well on similar tasks such as hate speech detection (Van Thin et al., 2019). The resulting model consists of four estimators, as illustrated in Figure 1.

The first set of classifiers are a multinomial naive Bayes classifier and a separate Gradient boosting classifier with tf-idf transformed vectors of the text transcriptions found in the data set. Originally we considered training the models on n-gram features, as previous research has shown that they can be useful to classify short texts (Buda and Bolonyai, 2020), but tf-idf vectors consistently performed better. Similarly, we experimented with adding more classifiers trained on tf-idf vectors, however, performance stayed consistent or decreased, thus we stayed with two.

Secondly, we introduce a Random forest classifier trained on a variety of features pertaining to the images themselves, rather than the text. The features used are Hu moment invariants (Hu, 1962), Haralick textures (Haralick et al., 1973) and image histograms. All three features are expected to help in gaining general insights into the image structure of the meme: Hu moment invariants are used to characterize the shape of an object in an image, Haralick textures should provide information about regions of interest, and the image histogram are employed to gain information about the color distribution, as the former two features require images to be converted into grayscale. To make all images equivalent, they were re-scaled to 500x500 pixels.

Lastly, with this being a multi-modal classifica-

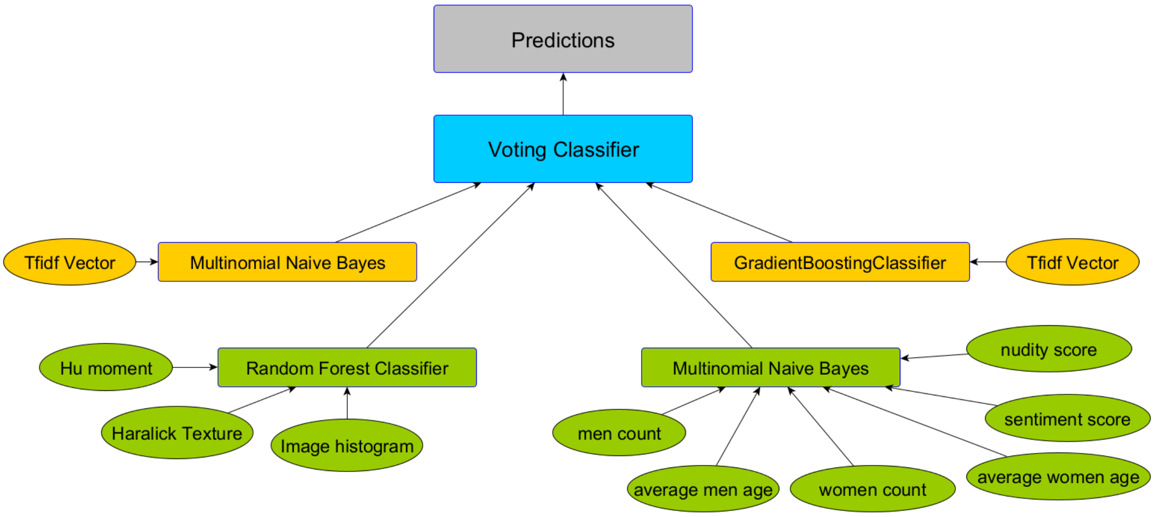

Figure 1: Ensemble Model used for Subtask A.

tion task, we were interested in whether we can extract additional information from the images/text transcriptions that might be relevant to decide if a meme is misogynistic or not. Thus, we enhanced the data set with the following features:

- The number of men/women depicted in the meme, as there might be a gender imbalance relevant for the classification process in combination with other features. Examining the text transcriptions, 'women' and 'woman' were the first and third most used word in misogynistic memes. It is not unreasonable to assume that women are also more likely to be featured in the corresponding images.
- For the same reason as above, we included the average estimated age of all men/women depicted in the meme.
- A binary nudity score, indicating whether an image is sexually explicit or not. In the training data, 15.75% of the misogynistic and only 2.86% of the non-misogynistic memes were sexually explicit.
- A sentiment score ranging from [-1, 1], indicating the polarity of any given text. A score of 1 indicates a positive statement, a score of -1 a negative sentiment. Misogynistic memes are usually hateful, so it can be assumed that the are more likel to use negative language in their texts.

All numerical scores retrieved from these features were transformed into a vector representation and used to train another multinomial naive Bayes classifier.

Finally, we built an ensemble model, using a majority voting rule, with all estimators created so far. Said model was hyperparameter tuned using a grid search, as well as cross-validated using five folds.

3.2 Experimental Setup

To extract information about the number of men/women and their estimated age, we made use of the Facial Recognition API provided by Face++.[2] The nudity score was obtained from the images using the NudeClassifier from NudeNet.[3] The classifier returns probabilities whether an image is sexually explicit or not. These probabilities were then transformed into a binary label. The sentiment was obtained using the TextBlob[4] library and the polarity scores added to the data set. The image features for the Random forest classifier were calculated through the opencv-python library,[5] whereas the tf-idf vectors, as well as the model itself, was built with the scikit-sklearn library (Pedregosa et al., 2011).[6]

The model was hypertuned using a grid search and a 5-fold cross-validation. The parameters for the grid search can be found in Table 2. The tf-idf vectors were tuned separately for the Multinomi-

[2] https://www.faceplusplus.com/
[3] https://github.com/notAI-tech/NudeNet
[4] https://github.com/sloria/textblob
[5] https://github.com/opencv/opencv-python
[6] https://scikit-learn.org/stable/index.html

nalNB classifier and the GradientBoostingClassifier. Similarly, the two MultinominalNB classifier were tuned separately, one in combination with the tf-idf vectors, and on in combination with the gender, age, nudity, and sentiment features.

Grid search parameter settings	
TfidfVectorizer	
ngram_range	(1,1), ***(1,2)***, (1,3), (2,2), (2,3)
analyzer	char, ***word***
max_features	***None***, 5000, 10000
MultinominalNB	
alpha	0.5, ***1.0***, 3.0
fit_prior	True, ***False***
RandomForestClassifier	
n_estimators	100, 1000, 5000, ***7000***
GradientBoostingClassifier	
n_estimators	100, 1000, 5000, ***7000***

Table 2: Grid search parameters for ensemble model. The final parameters are highlighted.

3.3 Results

Model	**Macro F1**
(1) MultinominalNB (tf-idf)	0.626
(2) GradientBoosting + (1)	0.645 *(+0.019)*
(3) MultinomialNB (gender) + (2)	0.649 *(+0.004)*
(4) RandomForest + (3)	**0.665** *(+0.016)*

Table 3: Gradual built-up of the ensemble model.

Our system for Subtask A achieved place 5 on the leaderboard, with a macro F1 score of 0.665. To make sure that each estimator of our ensemble model actually improves the classification of misogynistic memes, we gradually built it up and run it against the test data. As can be seen from the results in Table 3, both the GradientBoostingClassifier as well as the RandomForestClassifier provide substantial performance gains compared to the handcrafted gender, age, nudity, and sentiment features.

While the nudity score seems to offer strong support on whether a meme is misogynistic or not, the same cannot be said for the other handcrafted features. Given the sentiment score, for example, 1119 out of 5000 misogynistic memes received a negative sentiment score. Compared to 1103 non-misogynistic memes that also received a negative sentiment score, it seems likely that the score provides very little information for the classification process. Similarly, a direct correlation cannot be derived from the number of men/women or their estimated average age. Because of that, it is likely that they provide little information gain as well. The ensemble model from Subtask A was able to correctly classify instances of more explicit forms of misogyny. Memes that include the word “rape”, mention “cooking” or “cleaning”, show (exposed) female body parts in the image or explicitly mention them in the text are correctly identified as sexist. When only one of these features is present, the model produces false positives, for example harmless memes that feature women doing housework. The memes the model was not able to identify as misogynistic often do not feature human faces or only represent them by minimalistic drawings, are low quality, or have more subtle references to sexuality that the nudity detection cannot identify (such as handprints on breasts).

4 Subtask B: Multi-label classification of misogyny types

Although the ensemble model performed reasonably well for Subtask A, performance was significantly worse when using it for the extended misogynistic labels introduced in Subtask B. Because of this, we built and trained a different model infrastructure using a deep learning approach (which performed worse than our ensemble model, when applied to Subtask A).

4.1 System Overview

Subtask B is a multi-label classification task, so each meme can be assigned to one or more categories. To capture the features which differentiate those memes among different subtypes, we made use of a multi-label model from the Simple Transformers library,[7] using text data only.

In this task, we have five distinct binary labels for each data entry in the training set: misogyny, shaming, stereotype, objectification, and violence. The transformer-based model consists of a transformer model plus a classification layer on top of it. The main difference between this model and the binary classification model is that, in this model, the classification layer has five output neurons, corresponding to each out of the five labels in the training set. For the transformer model, we chose a pretrained RoBERTa model, which is imported by the Simple

[7] https://simpletransformers.ai/

	Misogyny (model A)		Shaming		Stereotype		Objectification		Violence	
	1 pred	0 pred	1 pred	0 pred	1 pred	0 pred	1 pred	0 pred	1 pred	0 pred
1 true	**333**	167	54	92	160	**190**	148	**200**	58	**95**
0 true	168	**332**	**98**	**756**	137	**513**	122	**530**	32	**815**

Table 4: Confusion matrices for all classification categories. Analysis was performed on the test set.

Transformer library from the Transformer library (Wolf et al., 2020) that was developed by HuggingFace. It is based on the RoBERTa model proposed by Liu et al. (2019).

RoBERTa removed the next sentence prediction task from BERT (Liu et al., 2019) which is one of the reasons why we chose RoBERTa over BERT, as the majority of the text transcriptions of memes in this task are consisting of either one or two sentences. The other reason being, that RoBERTa was trained on a larger dataset than BERT, thus more likely to result in better predictions.

4.2 Experimental Setup

The SimpleTransformers library is designed with the purpose of easily setting up a transformer model. Transforming text into a suitable vector representation is done automatically, and various hyperparameters can be tuned to improve performance. we set the threshold in our implementation to 0.8. The model was trained for 20 epochs using a GPU provided by the Kaggle notebook environment.[8]

4.3 Results

Our system achieved place 19 on the leaderboard, with a macro F1 score of 0.637 using the above-mentioned transformer-based model. We have also experimented with other approaches like the FastText library[9] which did not show better performance on this multi-label classification task than the transformer approach.

Compared to the true and false negatives in the ensemble model's prediction from Subtask A, which are very balanced, the model for Subtask B produces a lot more false negatives for the test set(see Table 4). True and false positives and negatives are very balanced in the ensemble model's predictions from Subtask A, as shown in Table 4. The model for Subtask B on the other hand produces more false negatives than positives for all categories except shaming. For this task, it would be favorable if the model was stricter and produced more false positives instead. The classifier performed reasonably well, even without information about the images. Still, some memes, especially in the stereotype and objectification category, cannot be understood to belong in that category without this information. Only about a third of the pictures (379 of 1000) are completely correctly identified when it comes to the four categorization labels, but 731 have at least 3 labels matching. Only 9 samples in the test set were classified in a way that no label matched the gold standard and 67 matched less than 2.

5 Conclusion

In our experiments, we explored and compared different models for multi-modal analysis of misogynistic memes. Surprisingly, ngram-models on both word and character levels did not perform as well as expected for this task. We discovered that ensemble models using text and image information can work well even if the text classifier uses simple features such as tf-idf vectors. BERT based text analysis performs better than the baselines, even if image features are not included. We found that an ensemble model can be improved through gradient boosting and adding information about nudity, age, gender of depicted humans, and text sentiment. In the future, it would be worth exploring how well a similar ensemble model with an (additional) BERT-classifier or other more powerful text classifier performs. For Subtask B, a RoBERTa-based multi-label classification model showed its power with purely text information. It would be interesting to train it with a class of weights and different threshold values. However, we were not able to create a well-performing multi-modal model that uses image information, which might be another interesting direction for further studies.

[8] https://www.kaggle.com/docs/notebooks

[9] https://fasttext.cc//

References

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on Twitter. In *International Conference on Applications of Natural Language to Information Systems*, pages 57–64. Springer.

Jakab Buda and Flóra Bolonyai. 2020. An ensemble model using n-grams and statistical features to identify fake news spreaders on Twitter Notebook for PAN at CLEF 2020.

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*.

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.

Simona Frenda, Bilal Ghanem, Manuel Montes-y Gómez, and Paolo Rosso. 2019. Online hate speech against women: Automatic identification of misogyny and sexism on Twitter. *Journal of Intelligent & Fuzzy Systems*, 36(5):4743–4752.

Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein. 1973. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621.

Ming-Kuei Hu. 1962. Visual pattern recognition by moment invariants. *IRE transactions on information theory*, 8(2):179–187.

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.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pranaydeep Singh, Nina Bauwelinck, and Els Lefever. 2020. LT3 at SemEval-2020 task 8: Multi-modal multi-task learning for memotion analysis. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1155–1162, Barcelona (online). International Committee for Computational Linguistics.

Shardul Suryawanshi and Bharathi Raja Chakravarthi. 2021. Findings of the shared task on troll meme classification in Tamil. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 126–132, Kyiv. Association for Computational Linguistics.

Dang Van Thin, Lac Si Le, and Ngan Luu-Thuy Nguyen. 2019. NLP@UIT: Exploring feature engineer and ensemble model for hate speech detection at VLSP 2019. *Training*, 5:3–51.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.