I2C at SemEval-2022 Task 5: Identification of Misogyny in Internet Memes

Pablo Cordón Hidalgo
Escuela Técnica Superior de Ingeniería
Universidad de Huelva (Spain)
pablo.cordon113@alu.uhu.es

Pablo Gonzalez Díaz
Escuela Técnica Superior de Ingeniería
Universidad de Huelva (Spain)
pablo.gonzalez682@alu.uhu.es

Jacinto Mata Vázquez
Escuela Técnica Superior de Ingeniería
Universidad de Huelva (Spain)
mata@uhu.es

Victoria Pachón Álvarez
Escuela Técnica Superior de Ingeniería
Universidad de Huelva (Spain)
vpachon@uhu.es

Abstract

In this paper we present our approach and system description on *Task 5 A in MAMI: Multimedia Automatic Misogyny Identification.* In our experiments we compared several architectures based on deep learning algorithms with various other approaches to binary classification using Transformers, combined with a nudity image detection algorithm to provide better results. With this approach, we achieved a test accuracy of 0.665.

1 Introduction

Misogyny is hatred or contempt for women. It is a form of sexism used to keep women at a lower social status than men, thus maintaining the societal roles of patriarchy. Misogyny has been widely practiced for thousands of years. It is reflected in art, literature, human societal structures, historical events, mythology, philosophy, and religion worldwide (Manne, 2017).

The Internet represents for many an extension of our offline interactions, and seemingly mundane everyday practices (e.g. participating in social media) form a significant part of our everyday experiences. Unfortunately, it is too common to see examples of harassment towards women and marginalized groups online within these experiences and practices (Drakett et al., 2018).

Women have a solid presence on the web, especially in picture-based web media like Twitter and Instagram: 78% of females utilize online media on numerous occasions each day, in contrast with 65% of men.

A popular way of communicating via social media platforms are MEMEs. A meme is an image portrayed through pictorial content with overlaid text which is written a posteriori, with the fundamental objective of being entertaining and/or ironic. Even though most of memes are created with the goal of making amusing jokes, shortly after their standardization individuals began to use them to disseminate hate against women, leading to sexist and aggressive messages in internet environments that allow people to freely express sexism without the fear of retaliation.

The detection of this disrespectful content is essential to eliminate it as soon as possible and stop spreading misogyny as a “joke”.

In MAMI: Task 5, Track A (meme binary classification) (Fersini et al., 2022), participants must determine whether a meme (text + image) is misogynist or not.

Meme sentiment-related tasks analysis is challenging, as memes are used created for various purposes, they are always evolving and often use sarcasm and humour. While misogyny and hate speech detection in text has been widely explored by the NLP community (Badjatiya et al., 2017). Its detection in images and text and how they correlate has not been explored in depth.

In this field we used BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2021) for training and classifying text, and nudenet tool for image classification (https://github.com/notAI-tech/NudeNet).

Our results show that combining text and image classification results are slightly better than using only one of the two methods.

The rest of the paper is organized as follows. Section 2 contains a briefly description of the dataset and its structure, Section 3 features the analysis of some of the previous works related to our task. In section 4 we describe the different

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 689 - 694

models and algorithm used, and their configurations. Section 5 provides details about data and models setup, while Section 6 reports experimental results and the paper is concluded in Section 7.

2 Background

This paper is focused on subtask A: Binary classification. The corpus provided is composed of 5000 1-value rows (misogyny) and 5000 0-value rows (not misogyny), so it is already well-balanced. Each row has the following format:

file_name | misogynous | shaming | stereotype | objectification | violence | text transcription

Where "file_name" is the .jpg link to the image, and "text transcription" is the text extracted from the image.

For our binary classification task, we only need the fields "file_name", "misogynous" and "text transcription" to train the models presented. For models that also used nudity recognition, a column "unsafe" was added later.

The csv file used follows this structure:

- file_name: *"10.jpg"*
- misogynous: "*1*"
- text transcription: "*ROSES ARE RED, VIOLETS ARE BLUE IF YOU DON'T SAY YES, I'LL JUST RAPE YOU quickmeme.com*"

3 Related work

Sentiment analysis of text is a very active research area that still faces multiple challenges such as irony and humour detection (Farías et al., 2016). In this area, the focus of the NLP community has increased towards detection of offensive language, aggression, hate-speech detection (Wei et al., 2021) and specifically misogyny, taking into account it can be expressed in a direct, explicit manner or an indirect, sarcastic manner, and even if this message is genered or not-gendered (Samghabadi et al., 2020).

The analysis of misogyny in memes has already been done in a psychologic point of view (Drakett et al., 2018), but never in computational models. Multimodal analysis research has been extended during the last years, but the focus was mostly on Video and text or speech and text (Pozzi et al., 2016). The specific multi-modality of memes in sentiment analysis has only been addressed recently by investigating their correlation with other comments in online discussions (French, 2017).

The growing usage of memes as an alternative medium of communication on social media has also recently drawn the attention of the online abuse research community.

However, memes completely make sense only if one takes both text and image content into account. These modalities can also lead to totally different perceived sentiment when recombined. For example, a meme whose image is a scary clown and the text is "happy birthday" will have a very different sentiment from a meme with the same text but with an image of a funny clown.

Sabat et al. (2019) performed hate speech detection on memes and showed that images were more important than text for the prediction.

In the other hand, Bonheme and Grzes (2020), investigated the relationship of text and image in sentiment analysis of memes, and found that images and text were uncorrelated. Fusion-based strategies did not show significant improvements and using one modality only (text or image) tends to lead to better results.

4 System overview

We focus on exploring different training techniques for text using BERT and RoBERTa, given their superior performance on a wide range of NLP tasks, while for image we used the python module *nudenet*.

Each text encoder, image classifier and training method used in our model are detailed below.

4.1 Text Encoders

BERT (Devlin et al., 2018): pretrained model BERT-base uncased, released by the authors, was used as embedding layer, tokenizer and classifier. It consists of 12 transformer layers, 12 self-attention heads per layer, and a hidden size of 768.

RoBERTa (Liu et al., 2021): We use the RoBERTa-base model released by the authors. Like BERT, RoBERTa-base consists of 12 transformer layers, 12 self-attention heads per layer, and a hidden size of 768.

4.2 Image nudity classification

As memes are mostly done as "joke" and tend to be ironical and use a very refined and deep text-image relationship.

A similar approach to the one used by Messina et al. (2021) was applied, where one of the two modalities acts as the main one and the second intervenes to enrich the first (in our case, text will act as the main modality and image will be used just to enrich the results from the first one).

The goal of this task was to detect misogyny, and we decided to use a Not Safe For Work (NSFW) image classifier.

Nudenet is the classifier used for this task. It gives each image of our dataset an "unsafe" value from 0 (safe) to 1 (unsafe). The Neural Net for Nudity Classification is trained on 160,000 entirely auto-labelled (using classification heat maps and various other hybrid techniques) images.

A NSFW image classifier was used for two main reasons. First, because image-only classification using Convolutional Neural Networks did not reached good results using the training data. We only obtained an accuracy of 0.52. Second, because most of NSFW images are misogynous, as it could be demonstrated by using only the *Nudenet* classifier, obtaining a value of 0.83 for precision in the positive (misogynous) class.

4.3 Models

Based on Convolutional Neural Networks (Konda et al., 2019), BiLSTM Neural Networks (Zhou et al., 2016), and BERT Transformers (Devlin et al., 2018), several models have been developed: (1) BiLSTM Neural Network with RoBERTa as embedding, (2) BiLSTM Neural Network with BERT as embedding, (3) BiLSTM Neural Network with BERT as embedding and nude detection, (4) 1- Dimensional Convolutional Neural Network with BERT Tokenizer, and (5) BERT Transformer with nude detection.

In our models, RoBERTa and BERT were used with word-embedding strategies, as they have an advantage over models like Word2Vec. Each word, under Word2Vec, has a fixed representation regardless of the context within which the word appears. Nevertheless, BERT produces word representations that are dynamically informed by the words around them (Shi and Lin 2019a).

For example, given the two sentences *"The man was director of a bank in his hometown."* and *"The man went fishing by the bank of the river."*, Word2Vec would produce the same word embedding for the word "bank" in both sentences. However, BERT will create different word embedding for "bank" for each sentence.

4.3.1 BiLSTM Neural Network with RoBERTa as embedding

Long Short-Term Memory (LSTM), (Hochreiter and Schmidhuber, 1997) is a widely known recurrent neural network (RNN) architecture. We used Bidirectional LSTM (Schuster and Paliwal, 1997) models for our experiments. A Bidirectional LSTM (BiLSTM) layer processes the text both in the forward as well as backward direction and hence is known to provide better context understanding.

Introduced by Facebook, the Robustly optimized BERT approach RoBERTa, is a retraining of BERT with an improved training methodology, 1000% more data and compute power (Shi and Lin, 2019b).

The RoBERTa model used to extract the word embedding layer for the BiLSTM Neural Network was RoBERTa-base uncased.

4.3.2 BiLSTM Neural Network with BERT as embedding

For this approach, a BERT-based model was used. In particular, we implemented the BERT-base uncased model.

4.3.3 BiLSTM Neural Network with BERT as embedding + nude detection

Python's Nudenet module was used in every image of the given dataset to assign it a value of "unsafety" from 0 (safe) to 1 (unsafe). Then, if the text prediction is 1, the final prediction is set as 1, otherwise if the text prediction is 0 and the image unsafe value is greater than a threshold value, final prediction is set to 1.

The best values for this threshold were the ones with the best accuracy classifying using with nudity in images. These threshold values were 0.45 and 0.60.

4.3.4 1-Dimensional Convolutional Neural Network with BERT as embedding + nude detection

Convolutional neural networks (CNN) (Lecun et al., 1998) are originally designed to process and learn information from image features by applying convolution kernels and pooling techniques which are widely adopted for extracting stationary features; for instance, CNN has shown its adaptability in the field of text mining and NLP tasks. (Kim, 2014) reported series of experiments

with CNNs that achieve good results on sentence classification and sentiment analysis tasks.

To improve this model classification, BERT-base uncased model was used to create the word embedding layer. The nudity classification algorithm with a threshold of 0.45. was used to achieve better results.

4.3.5 BERT Transformer with nude detection

Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based machine learning technique for natural language processing (NLP) pre-training developed by Devlin et al. (2018). BERT model is pre-trained from unlabeled data extracted from the BooksCorpus with 800M words and English Wikipedia with 2,500M words.

It uses Transformer, (Vaswani et al., 2017) an attention mechanism that learns contextual relations between words (or sub-words) in a text. In its vanilla form, Transformer includes two separate mechanisms — an encoder that reads the text input and a decoder that produces a prediction for the task. Since BERT's goal is to generate a language model, only the encoder mechanism is necessary.

Also, the nudity classification algorithm with a threshold of 0.45. was used to achieve better results.

5 Experimental setup

For each text transcription row in the corpus, a small preprocessing was applied. Every word was undercased, web pages' links, hashtags, usernames, emojis, punctuation symbols, numbers, and words with length less than 2 characters were removed, as well as English stop words using nltk.corpus module (Sarica, 2021).

For all the experiments, we split the training dataset in two parts: 80% for training and 20% for validation using a stratify approach.

The parameters used in the training phase were: batch size of 32 and 5 epochs.

6 Results

Table 1 shows a summarization of the training and test results obtained for each one of the models in the evaluation phase.

According to the official metrics (F1-score for the positive class), our results are all around 0.60 – 0.65, being the best model BiLSTM Neural Network with BERT as embedding + nude detection, with a 0.665 F1-score. With this result, we obtained the 43 place in the ranking.

As we expected, the model that obtained the best results during the training phase also obtained the best result in the evaluation phase.

7 Conclusions

In this paper, several approaches and systems descriptions on Task 5 (Subtask A) in SemEval 2022: Multimedia Automatic Misogyny Identification are detailed. The main aim was to develop various deep learning models and check how multi-modality of text and image could help achieve better classification results.

Six different models were developed and BiLSTM Neural Network with BERT as embedding + nude detection (0.45 threshold) was the definitive one. After training and analyzing each model, we achieved an F1-score of 0.665 in

Model	**Training phase**		**Evaluation phase**
	Accuracy	**F1 - Score**	**F1 - Score**
BiLSTM Neural Network with RoBERTa as embedding	0.793	0.799	0.607
BiLSTM Neural Network with BERT as embedding	0.810	0.832	0.652
BiLSTM Neural Network with BERT as embedding + nude detection threshold 0.45	**0.837**	**0.840**	**0.665**
BiLSTM Neural Network with BERT as embedding + nude detection threshold 0.6	0.835	0.838	0.663
1-D Convolutional Neural Network with BERT as embedding + nude detection	0.813	0.825	0.649
BERT Transformers + nude detection	0.806	0.812	0.639

Table 1: Results obtained with the different models

the evaluation phase for class "1". We can conclude that merging text and image classifiers improves the results in the task of misogyny detection in memes.

In future works, we intend to improve our image classifiers models. Also, we want to use other pretrained models based on transformers.

References

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. "Deep Learning for Hate Speech Detection in Tweets." Perth, Australia, International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3041021.3054223.

Lisa Bonheme and Marek Grzes. 2020. *SESAM at SemEval-2020 Task 8: Investigating the Relationship between Image and Text in Sentiment Analysis of Memes* International Committee for Computational Linguistics. doi:10.18653/v1/2020.semeval-1.102.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." *CoRR* abs/1810.04805.

Jessica Drakett, Bridgette Rickett, Katy Day, and Kate Milnes. 2018. "Old Jokes, New Media – Online Sexism and Constructions of Gender in Internet Memes." *Feminism Psychology* 28 (1): 109-127. https://journals.sagepub.com/doi/full/10.1177/0959353517727560.

Delia Farías, Viviana Patti, and Paolo Rosso. 2016. "Irony Detection in Twitter." *ACM Transactions on Internet Technology* 16 (3): 1-24. http://dl.acm.org/citation.cfm?id=2930663

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.

Jean H French. 2017. "Image-Based Memes as Sentiment Predictors.". doi:10.23919/i-Society.2017.8354676.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation* 9: 1735-1780. doi:10.1162/neco.1997.9.8.1735.

Yoon Kim. 2014. *Convolutional Neural Networks for Sentence Classification*.

Srinivas Konda, B. Rani, Varaprasad Mangu, G. Madhukar, and B. Ramana. 2019. "Convolution Neural Networks for Binary Classification." *Journal of Computational and Theoretical Nanoscience* 16: 4877-4882. doi:10.1166/jctn.2019.8399.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. 1998. "Gradient-Based Learning Applied to Document Recognition." *Proceedings of the IEEE* 86 (11): 2278-2324. doi:10.1109/5.726791.

Zhuang Liu, Wayne Lin, Ya Shi, and Jun Zhao. 2021. "A Robustly Optimized BERT Pre-Training Approach with Post-Training." In *Chinese Computational Linguistics*, 471-484. Cham: Springer International Publishing. https://library.biblioboard.com/viewer/822f5f9f-f7f4-11eb-926c-0a9b31268bf5.

Kate Manne. 2017. *Down Girl: The Logic of Misogyny*. New York: Oxford University Press. https://oxford.universitypressscholarship.com/10.1093/oso/9780190604981.001.0001/oso-9780190604981.

Nicola Messina, Fabrizio Falchi, Claudio Gennaro, and Giuseppe Amato. 2021. "AIMH at SemEval-2021 Task 6: Multimodal Classification using an Ensemble of Transformer Models. "Association for Computational Linguistics https://aclanthology.org/2021.semeval-1.140.

F. Pozzi, E. Fersini, E. Messina, and B. Liu. 2016. *Sentiment Analysis in Social Networks* Elsevier Science. https://books.google.es/books?id=aS2lCgAAQBAJ.

Benet Oriol Sabat, Cristian Canton Ferrer, and Xavier Giro-i-Nieto. 2019. "Hate Speech in Pixels: Detection of Offensive Memes Towards Automatic Moderation." *arXiv Preprint arXiv:1910.02334*.

Niloofar Safi Samghabadi, Parth Patwa, Srinivas PYKL, Prerana Mukherjee, Amitava Das, and Thamar Solorio. 2020. "Aggression and Misogyny Detection using BERT: A Multi-

Task Approach." European Language Resources Association, may. https://aclanthology.org/2020.trac-1.20.

Serhad Sarica and Luo,Jianxi. 2021. "Stopwords in Technical Language Processing." *Plos One* 16 (8): 1-13. https://doi.org/10.1371/journal.pone.0254937.

Mike Schuster and Kuldip Paliwal. 1997. "Bidirectional Recurrent Neural Networks." *Signal Processing, IEEE Transactions On* 45: 2673. doi:10.1109/78.650093.

Peng Shi and Jimmy Lin. 2019a. *Simple BERT Models for Relation Extraction and Semantic Role Labeling* https://arxiv.org/abs/1904.05255.

———. 2019b. "Simple BERT Models for Relation Extraction and Semantic Role Labeling." http://arxiv.org/abs/1904.05255.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. *Attention is all You Need*.

Bencheng Wei, Jason Li, Ajay Gupta, Hafiza Umair, Atsu Vovor, and Natalie Durzynski. 2021. "Offensive Language and Hate Speech Detection with Deep Learning and Transfer Learning." . https://arxiv.org/abs/2108.03305.

Peng Zhou, Zhenyu Qi, Suncong Zheng, Jiaming Xu, Hongyun Bao, and Bo Xu. 2016. *Text Classification Improved by Integrating Bidirectional LSTM with Two-Dimensional Max Pooling*.