IITR CodeBusters at SemEval-2022 Task 5: Misogyny Identification using Transformers

Gagan Sharma
Indian Insitute of Technology Roorkee
gagan_s@cs.iitr.ac.in

Gajanan Sunil Gitte*
Indian Insitute of Technology Roorkee
gajanan_s@cs.iitr.ac.in

Shlok Goyal*
Indian Insitute of Technology Roorkee
shlok_g@cs.iitr.ac.in

Raksha Sharma
Indian Insitute of Technology Roorkee
raksha.sharma@cs.iitr.ac.in

Abstract

This paper presents our submission to task 5 (Multimedia Automatic Misogyny Identification) of the SemEval 2022 competition. The purpose of the task is to identify given memes as misogynistic or not and further label the type of misogyny involved. In this paper, we present our approach based on language processing tools. We embed meme texts using GloVe embeddings and classify misogyny using BERT model. Our model obtains an F1-score of 66.24% and 63.5% in misogyny classification and misogyny labels, respectively.

1 Introduction

SemEval 2022 task 5 (Fersini et al., 2022) is a sentiment analysis task aimed at memes[1], divided into two subtasks of increasing complexity. Subtask A is misogynous meme identification, *i.e.*, whether a meme should be categorized either as misogynous or not misogynous; subtask B, on the other hand, is a multi-label classification problem, *i.e.*, classifying what kind of misogyny is involved in the meme. A meme is usually intended to convey a sarcastic message, but in a short time, people have started to use them to deliver sexist messages in an online environment. The anonymity of the internet tends to make them more aggressive. Such an environment amplifies the offline world's sexual stereotyping and gender inequality.

Meme sentiment analysis is a challenging task as often it relies on implicit themes or knowledge and trending news at the time of the creation of the meme. Meme analysis has been of growing interest for the NLP community. Our approach also relies on various NLP-based tools and the code to implement the paper is available at https://github.com/gagansh7171/IITR-CodeBusters .

*These two authors contributed equally.

[1]The term meme used in this task refers to an idea or a message conveyed via an image and embedded text.

The rest of the paper is organized as follows. In Section 2 we briefly describe the details regarding the datasets used. In Section 3 we describe some recent related work conducted in the field of meme classification. In Section 4 we describe and define the models and baselines for the task. In section 5 we describe the low-level details including scoring parameters, pre-processing, libraries and hyper-parameters used for the experimental setup. In Section 6 we describe the results obtained and some insight of why the models are performing the way they performed. In section 7 we conclude the paper.

2 Background

The dataset used for this task consists of a set of 10000 memes images whose text has already been extracted. Each meme has already been labeled as *misogynous* for subtask A and as *shaming*, *stereotype*, *objectification* ,and *violence* for subtask B. The dataset is perfectly balanced in terms of classification of misogyny, *i.e.*, precisely 50% of the memes are misogynous and the rest 50% are not. Of these 5000 misogynous memes -

- 1274 or 25.48% are labelled as shaming category.
- 2810 or 56.2% are labelled as stereotype category.
- 2202 or 44.04% are labelled as objectification category.
- 953 or 19.06% are labelled as violence category.

The data is provided as zip file of memes in image format and a .csv file where the memes are labelled according to the following structure (Fersini et al., 2022).

- **file_name:** name of the file denoting the meme

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 728 - 732

- **misogynous:** a binary value (1/0) indicating if the meme is misogynous or not. A meme is misogynous if it conveys an offensive message having as target a woman or a group of women.
- **shaming:** a binary value (1/0) indicating if the meme is denoting shaming. A shaming meme aims at belittling women because of some body characteristics.
- **stereotype:** a binary value (1/0) indicating if the meme is denoting stereotype. A stereotyping meme aims at representing a fixed idea or preconceived notion regarding women.
- **objectification:** a binary value (1/0) indicating if the meme is denoting objectification. A meme that describes objectification represents a woman like an object through over-analysis of physical or sexual appeal or comparing women to inanimate objects.
- **violence:** a binary value (1/0) indicating if the meme is denoting violence. A violent meme describes physical or verbal violence or harassment represented by textual or visual content.
- **Text Transcription:** transcription of the text reported in the meme.

3 Related Work

Memes are essentially language of the internet but such widespread use of memes had also made them a target for spreading hateful and offensive messages by fringe web communities (Zannettou et al., 2018). Many online communities accept sexism and harassment in the name of humour in the form of memes and has resulted in increased attraction of attention from academics (Drakett et al., 2018). Thus, memes had emerged as a multi-modal expression of online hate.

Research in the field of multi-modal sentiment analysis has been mostly focused on video and text or speech and text (Rao et al., 2021; Zadeh et al., 2016). Sentiment analysis of memes was conducted by French (2017) but it was based on correlation of the meme and online discussion in the comments.

Recent study for meme classification was done by Zia et al. (2021) where hateful memes were classified based on the protected category they attacked which were *race*, *sex*, *religion*, *nationality*, *disability*. The study included usage of state-of-the-art visual and textual representations to produce respective embedding of the memes which were then concatenated to train a logistic regression classifier model. Facebook recently launched The Hateful Memes Challenge to accelerate development in this field (Facebook, 2020).

We follow a more humble approach of using GloVe to embed text in the memes but tried models more sophisticated than Logistic Regression Classifier for the classification. The study conducted by Zia et al. (2021) reflects that better results emerge when both text and image are considered. But we do not consider the images for the challenge and consider this challenge an opportunity to learn about NLP first-hand.

4 System Overview

BERT[2] model is a state-of-the-art model developed by the AI team at Google (Vaswani et al., 2017). Traditional NLP models are unidirectional, *i.e.*, they read the text from left-to-right or right-to-left. On the other hand, BERT is bidirectional, understanding the correlation of a word with words on both sides by reading the entire sequence of the words at once. Google showed that this scheme helps better understand the statement's sentiment, making this model the best fit for the task.

RoBERTa[2] model is built on top of the BERT model. It has the same architecture but differs in terms of tokenizer and pre-training scheme, *i.e.*, much larger mini-batches and learning rates are used. The architecture similarity with the BERT model makes this model suitable for this task.

Baseline scores are made available by the organizers and are mentioned in Table 1. We provide additional baseline scores obtained by traditional classification models as well for comparison with the BERT and RoBERTa model. The different baseline models used are Logistic Regression (LR)[3], K Nearest Neighbours (KNN)[3], Random Forest (RF)[3] and Multilayer Perceptron (MP)[3].

5 Experimental Setup

The scoring parameters, pre-processing, language, libraries and hyper-parameters used are mentioned in this section for ease in reproduction.

[2]We use transformers 4.16.2 implementation of the model.
[3]We use Scikit-learn 1.0.2 implementation of the model.

Model	A	B
Baseline_Image_Text	54.3%	0%
Baseline_Text	64%	0%
Baseline_Image	63.9%	0%
Baseline_Flat_Multilabel	43.7%	42.1%
Baseline_Hierarchical_M	65%	62%

Table 1: Baseline Scores obtained by SemEval Organizers.

5.1 Scoring parameters

Macro F1 score is used as a measure of performance for the models for subtask A and weighted average of F1 scores for each prediction category for subtask B.

5.2 Pre-processing

The text is converted to lowercase before being used for training. Each meme is vectorised using GloVe embedding (Pennington et al., 2014).

Glove consists of word to vectors mapping and these vectors come in various dimensions. We are using 200-d vectors. For embedding one meme we find out word-vectors for every word and calculate an average of these vectors to finally calculate a vector representation of a meme. Glove vectors map words to a point in n-dimensional space where words with similar meaning are closer to each other. So taking average of these vectors should give us a vector in this space which represent an average gist of the message.

5.3 Language and libraries used

The experiment is conducted at Google Colab platform using Python 3.7.2 as the programming language. The packages are listed in Table 2

Package	Version
Pandas	1.3.5
Jupyter	1.0.0
Keras	2.7.0
Tensorflow	2.7.0
Tensorflow-hub	0.12.0
Numpy	1.19.5
Transformers	4.16.2
Scikit-learn	1.0.2

Table 2: List of Python Packages used for the Experiment.

5.4 Hyper-parameters of the models

This section details hyper-parameters for each model.

K Nearest Neighbours 5 CPU jobs, other values are default from Scikit-learn (refer KNN).

Logistic Regression random seed of 42, solver is liblinear, maximum iterations of 1000, 5 CPU jobs, f1_macro scoring is used, refit is set to True, other values are default from Scikit-learn (refer LogisticRegressionCV).

Multilayer Perceptron maximum iteration is set to 200, other values are default from Scikit-learn (refer MLPClassifier).

Random Forest 5 CPU jobs, bootstrap samples are used while building trees, out-of-bag samples are used to estimate the generalization score, 10 trees are used in the forest, all features are considered while looking for best split, other values are default from Scikit-learn (refer RandomForestClassifier).

BERT transformers based implementation of BERT is used. Pretrained bert-base-uncased model and tokenizer are used. Adam optimizer with 3e-6 learning rate, 1e-08 epsilon and 1.0 clipnorm as parameters is used as the optimizer. Sparse Categorical Crossentropy with from_logits set to True is used as loss function. Early stopping is done with a patience value of 4 and minimum increment of validation accuracy as 0.005. The model which gave best result for validation set is restored before training is complete. Other values are default from Hugging-Face implementation (refer BERT).

RoBERTa transformers based implementation of RoBERTa is used. Pretrained roberta-base model and tokenizer are used. Rest of the parameters are same as used in BERT model. Other values are default from Hugging-Face implementation (refer RoBERTa).

6 Results

Models are trained with 80% of data as training data and 20% of data as validation data. The results obtained for subtask A and B are mentioned in Table 3 and 4 respectively. The models' hyperparameters were tuned for subtask A and were reused for subtask B, hence training and validation scores are omitted in Table 4, instead score for each classification category[4] and final scores are listed. Baseline results obtained by the SemEval

[4]These scores are obtained in post-evaluation phase after the labels for test data were released by the organizers.

Model	Training Score	Validation Score	Subtask A
KNN	78.05%	64.27%	57.54%
LR	73.95%	71.03%	58.69%
MP	89.97%	71.32%	58.05%
RF	98.56%	61.97%	58.81%
BERT	**87.26%**	**82.10%**	**66.24%**
RoBERTa	88.63%	80.60%	60.80%

Table 3: Scores obtained for subtask A.

Model	Misogynous	Shaming	Stereotype	Objectification	Violence	Subtask B
KNN	57.54%	53.62%	58.24%	54.51%	56.38%	56.50%
LR	58.69%	53.35%	55.63%	51.52%	52.65%	55.17%
MP	58.06%	52.30%	54.50%	60.66%	62.66%	57.74%
RF	58.88%	48.01%	54.19%	45.04%	47.16%	52.30%
BERT	**66.24%**	**64.06%**	**61.75%**	**65.20%**	**66.51%**	**64.76%**
RoBERTa	60.80%	46.06%	62.10%	66.11%	53.36%	60.14%

Table 4: Scores obtained for subtask B.

Organizers are listed in Table 1 for comparison.

From the results we can see that Random Forest model has high level of over-fitting as the score difference of training and validation is highest in this model.

The least difference in training and validation score appears in Logistic Regression model but this is mostly due to high level of under-fitting as the least training score is obtained in this model.
BERT model provides the best results. The model obtains a high training score and score difference in training and validation is the least for this model. The trend maintains in subtask B as well with the model obtaining highest score.

Ranking phase - We submit our scores obtained from BERT model for the leader board. The scores obtained are 66.24% and 63.5%[5] for subtask A and B respectively. The ranks obtained for subtask A and B are **47** and **34** respectively.

7 Conclusion

The results show that BERT model is the best among those we tried for the given problem statement. However, these results are obtained using purely NLP based tools and techniques, the image component of the memes is not considered. As discussed in the Background section, there is a scope of improvement if we consider the images of the memes as well for the classification.

[5]The score of 64.76% mentioned in the table is obtained using the same model used in making the submission for evaluation phase. This score is obtained post-evaluation with test-labels being made public by the organizers.

References

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.

Facebook. 2020. Hateful memes challenge and dataset for research on harmful multimodal content.

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. In *2017 International Conference on Information Society (i-Society)*, pages 80–85.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Ashwini Rao, Akriti Ahuja, Shyam Kansara, and Vrunda Patel. 2021. Sentiment analysis on user-generated video, audio and text. In *2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, pages 24–28.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88.

Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*, pages 188–202, New York, NY, USA.

Haris Zia, Ignacio Castro, and Gareth Tyson. 2021. Racist or sexist meme? classifying memes beyond hateful. pages 215–219.