

# iCompass Working Notes for Arabic Misogyny Identification

Abir Messaoudi<sup>1</sup>, Chayma Fourati<sup>1</sup>, Mayssa Kchaou<sup>1</sup> and Hatem Haddad<sup>1</sup>

<sup>1</sup>*iCompass, Tunisia*

## Abstract

We describe our submitted system to the first Arabic Misogyny Identification shared task. We tackled both subtasks, namely Misogyny Content Identification (Subtask 1) and Misogyny Behavior Identification (Subtask 2). We used state-of-the-art Machine Learning models and pretrained contextualized text representation models that we fine-tuned according to the downstream task in hand. As a first approach, we used Machine Learning algorithms including: Naive Bayes and Support Vector Machine for both subtasks. Then, we used Google's multilingual BERT and then other BERT Arabic variants: AraBERT, ARBERT and MARBERT. The results found show that MARBERT outperforms all of the previously mentioned models overall, whether on Subtask 1 or Subtask 2.

## Keywords

Misogyny, Machine Learning, BERT, Finetuning

## 1. Introduction

Nowadays, social media presents an important role in the spread of misogynistic behaviour. Hence, misogyny identification presents a trending task, particularly in Arabic since it has different variants and dialects across the world. Even if some dialects share some vocabulary, they still differ according to countries, where each dialect has its own specifications. Because of the massive amount of such content, automatic identification of misogynistic behaviours becomes crucial. The paper is structured as follows: Section 2 provides a concise description of the used dataset. Section 3 describes the used systems and the experimental setup to build models for Misogyny Content Identification and Misogyny Behavior Identification. Section 4 presents the obtained results. Finally, section 5 concludes and points to possible directions for future work.

## 2. Data

The provided train dataset [1] of the competition [2] consists of **7866 tweets** written in Modern Standard Arabic (MSA) and several Arabic dialects including: Gulf, Egyptian and Levantine. The dataset has two label columns: misogyny and category for the first and second subtasks respectively. The first subtask consists of a binary classification problem, where the column misogyny contains two labels (Misogyny and None). The second subtask consists of a multiclass


---

*FIRE 2021: Forum for Information Retrieval Evaluation, 13th-17th December, 2021*

✉ abir@icompass.digital (A. Messaoudi); chayma@icompass.digital (C. Fourati); mayssakchaou933@gmail.com (M. Kchaou); hatem@icompass.digital (H. Haddad)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

classification problem, where the column category contains eight labels as follows: none, damning, derailling, discredit, dominance, sexual harassment, stereotyping & objectification, and threat of violence. In order to validate our models, we split the provided train dataset into train, dev and test with ratios 70%, 10% and 20% respectively. Tables 1 and 2 present statistics of the splitted train dataset for Subtask 1 and Subtask 2 respectively.

**Table 1**

Splitted train dataset statistics for Subtask 1.

Class	Train	Validation	Test
None	2053	221	787
Misogyny	3609	409	787
<b>Total</b>	<b>5662</b>	<b>630</b>	<b>1574</b>

**Table 2**

Splitted train dataset statistics for Subtask 2.

Class	Train	Validation	Test
None	2069	205	787
Damning	524	44	101
derailing	88	8	9
discredit	2131	261	476
dominance	159	26	34
Sexual harassment	47	2	12
Stereotyping & objectification	473	59	121
Threat of violence	171	25	34
<b>Total</b>	<b>5662</b>	<b>630</b>	<b>1574</b>

Train, dev and test datasets were preprocessed by removing links (https, //, etc..) emoji symbols (:p, :D, etc..) , hashtags (#), tags (@) retweets (RT) and punctuation (?!, etc..). An example is the tweet "❤️ انتي بطلة @مستخدم". After preprocessing, the latest becomes "انتي بطلة".

### 2.1. Third Party Dataset for Training

At iCompass, we gathered our Tunisian Misogyny dataset labelled as None (0) and Misogyny (1) collected from Tunisian sources. We added this dataset for the training of the first subtask since it has the same labels. Hence, the dataset was enhanced by 818 tweets labelled as "None" and 642 labelled as "Misogyny". The same preprocessing techniques were performed. However, the new obtained dataset was not used for the experiments, but only when submitting our results.

## 3. System Description

As a first approach, we used two Machine Learning algorithms: Naive Bayes (NB) and Support Vector Machine (SVM) chosen based on the state of the art with a variation of hyperparameters in order to find the best performing values.

Pretrained contextualized text representation models have shown to perform effectively in order to make a natural language understandable by machines. Bidirectional Encoder Representations from Transformers (BERT) [3] is, nowadays, the state-of-the-art model for language understanding, outperforming previous models and opening new perspectives in the Natural Language Processing (NLP) field. Hence, as a second approach, we used multilingual cased BERT model (mBERT) [3] since it contains more than 100 languages including the Arabic one. Then, we used three BERT Arabic variants: AraBERT [4], ARBERT [5] and MARBERT [5].

After different experiments, MARBERT achieved the best results for the two subtasks: Misogyny Content Identification and Misogyny Behavior Identification. We believe this is because MARBERT was trained mostly on dialectal Arabic which was underrepresented in previous pretrained models. Since this task’s data is multi-dialectal, this model is expected to achieve the best performance.

We trained our models on a Google Cloud GPU of 8 cores using Google Colaboratory. The final models that we used to make the submissions are:

- For Misogyny Content Identification: a model based on MARBERT, trained for 4 epochs with a learning rate of  $2e-5$ , a batch size of 32 and max sequence length of 128.
- For Misogyny Behavior Identification: a model based on MARBERT, trained for 4 epochs with a learning rate of  $2e-5$ , a batch size of 32 and max sequence length of 128.

## 4. Results and Discussion

We submitted two runs to each subtask: run1 is trained on the provided train dataset, and run2 on the augmented train dataset.

### 4.1. Sub-task A - Misogyny Content Identification

This subtask is a binary classification problem which includes labels "None" and "Misogyny". Table 4 presents the results of experiments performed for this subtask where the best result was achieved by MARBERT.

**Table 3**  
Results obtained for Subtask 1.

Model	Accuracy	F1 macro	F1 micro
SVM	79%	79%	79%
NB	80%	79%	79%
MBERT	75%	74%	74%
ARABERT	81%	81%	81%
ARBERT	80%	80%	80%
<b>MARBERT</b>	<b>89%</b>	<b>89%</b>	<b>89%</b>

### 4.2. Sub-task B - Misogyny Behavior Identification

This subtask is a multiclass classification problem, including eight labels. Table 4 presents the results of experiments performed for this subtask where the best result was also achieved by

MARBERT. Because the dataset is not balanced, F1 macro gives low performances.

**Table 4**

Results obtained for Subtask 2.

Model	Accuracy	F1 macro	F1 micro
SVM	71%	32%	69%
NB	70%	41%	70%
MBERT	66%	30%	64%
ARABERT	70%	32%	68%
ARBERT	63%	24%	59%
<b>MARBERT</b>	<b>83%</b>	<b>52%</b>	<b>82%</b>

### 4.3. Official Submission Results

The results obtained on the final released test dataset are presented in table 5.

**Table 5**

Results on the final test datasets.

Subtask	Accuracy	Precision	Recall	F1 score
Subtask 1 run 1	83.3%	82.6%	82%	82.3%
Subtask 1 run 2	50.8%	50.2%	50.3%	49.9%
Subtask 2 run 1	63.7%	24.2%	24.8%	24.5%
Subtask 2 run 2	63.7%	24.2%	24.8%	24.5%

The augmented train dataset contains comments in the Tunisian dialect, which may have led to a decrease in the results of the first Subtask. Hence, run 1 outperforms run 2 in the subtask 1. Results of Subtask 2 are the same because we did not increase the train dataset with our Tunisian Misogyny one.

## 5. Conclusion

In this work, two Machine Learning (SVM and NB) and four language models were used to classify misogyny and to detect misogynic behaviour (mBERT, AraBERT, ARBERT and MARBERT). The best results were obtained by MARBERT for both tasks with different hyperparameters, which was selected for the final submission. Future work would involve working on bigger contextualized pretrained models and enriching the existing Misogyny Content and Misogyny Behaviour datasets.

## References

- [1] H. Mulki, B. Ghanem, Let-Mi: An Arabic Levantine Twitter Dataset for Misogynistic Language, in: Proceedings of the 6th Arabic Natural Language Processing Workshop (WANLP 2021), 2021.

- [2] H. Mulki, B. Ghanem, ArMI at FIRE2021: Overview of the First Shared Task on Arabic Misogyny Identification, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.
- [3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.
- [4] W. Antoun, F. Baly, H. Hajj, AraBERT: Transformer-based model for Arabic language understanding, in: Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, 2020, pp. 9–15.
- [5] M. Abdul-Mageed, A. Elmadany, E. M. B. Nagoudi, Arbert & marbert: Deep bidirectional transformers for arabic, ArXiv abs/2101.01785 (2021).