Context-aware Language Modeling for Arabic Misogyny Identification

Istabrak Abbes[1], Eya Nakache[1] and Moez BenHajHmida[1]

[1]*University of Tunis El Manar, Campus Universitaire El Manar, Le Belvedere, 1002, Tunisia.*

Abstract

In this paper, we describe our efforts on the shared task of Arabic Misogyny Identification (ArMI) [1]. We tackled the Misogyny Content Identification subtask (Subtask-1). Our experiments were based on preprocessing the given data, then fine-tuning pretrained MARBERT language model on the Misogyny Identification downstream task. Experimental results performed only on Subtask-1, show that keeping emojis in text can influence the model.

Keywords

Natural Language Processing, classification, BERT, MarBERT, misogyny, Arabic Dialects

1. Introduction

Most Arabic interactions in media (TV, radio, etc), and on the internet (social media, forums) are produced in local dialects. Dialectal Arabic (DA) is significantly different from the formal Arabic language, known as Modern Standard Arabic (MSA). Especially on social media, we observe various dialects and free writing forms that make the Natural Language Processing task more complicated.

Misogyny, which is defined as the hate towards women, or the notion that men are far superior to women, has spread across a range of social media platforms, becoming a global epidemic. Women in the Arab world face a wide range of online misogyny, which sadly reinforces and excuses gender inequality, violence against women, and women's undervaluation. From here came the challenge of misogyny detection in Arabic dialects.

Therefore, the Arabic Misogyny Identification (ArMI) task is the first shared task that attempts to address the issue of automatic identification of Arabic online misogyny. The ArMI shared task attempts to identify the misogynistic content and recognize the different misogynistic behaviors in a collection of Arabic (MSA/dialectal) tweets [1].

In this paper, we describe the system submitted for the ArMI shared task on misogyny detection in Arabic dialects. In this challenge, we conducted experiments on MarBERT [2], a BERT-based [3] model that focused on both Dialectal Arabic (DA) and MSA. We utilized the MarBERT pretrained model that we fine-tuned on the provided training set after applying our

Forum for Information Retrieval Evaluation, December 13-17, 2021, India

istabrak.abbes@etudiant-enit.utm.tn (I. Abbes); eya.nakache@etudiant-enit.utm.tn (E. Nakache); moez.benhajhmida@enit.utm.tn (M. BenHajHmida)

0000-0002-9421-8566 (M. BenHajHmida)

preprocessing strategy.

The rest of the paper is organized as follows: in Section 2 we introduce ArMI dataset. In Section 3, we describe our approach in tackling the problem. Then, in Section 4 we provide and discuss the results of the proposed method on Subtask-1. Section 5 concludes our work throughout this shared task.

2. Data

The released train dataset [4] for the ArMi competition [1] is the same for both shared subtasks. The organizers provided a dataset containing 7866 tweets for training and tweets for testing. The dataset was annotated for misogyny detection task (Subtask-1) with the label "misogyny" for misogynistic tweets and "none" for non misogynistic tweets. Table 1 presents statistics of the train dataset for Subtask-1. For the second shared task (Subtask-2) on Misogyny Behavior Identification the labels are (Damning, Derailing, Discredit, Dominance, Sexual Harassment, Stereotyping and Objectification, Threat of Violence or None) for respectively:

- Damning (Damn): tweets under this class contain cursing content.
- Derailing (Der): tweets under this class combine justification of women abuse or mistreatment.
- Discredit (Disc): tweets under this class bear slurs and offensive language against women.
- Dominance (Dom): tweets under this class imply the superiority of men over women.
- Sexual Harassment (Harass): tweets under this class describe sexual advances and sexual nature abuse.
- Stereotyping and Objectification (Obj): tweets under this class promote a fixed image of women or describe women's physical appeal.
- Threat of Violence (Vio): tweets under this class have an intimidating content with threats of physical violence.
- None: if no misogynistic behaviors exist.

Table 1
Train dataset statistics for Subtask-1.

Class	Number of samples
None	3061
Misogyny	4805
Total	7866

3. System

This section describes the various data preparation procedures and models utilized in the experiments.

3.1. Preprocessing

To prepare the dataset for preprocessing, we completed 4 main stages as follows:

- **Cleaning:** we removed all of the diacritics such as (damma,tashdid, fatha, kasra, etc.), English words and numbers, English and Arabic punctuations, URLS and USER mention tokens.
- **Elongation removal:** any repeated character for more than twice was removed. For example, the word "اكيدددد " becomes "اكيد" after the preprocessing.
- **Letter normalisation:** Arabic characters that appeared in a variety of forms were combined into a single form. For example, letter like آ , إ, and أ are replaced with a ا
- **Extract hashtag keywords:** To extract intelligible key phrases, we deleted the hash symbol "#" and replaced the underscore "_" within a hashtag with a white space. For exemple " #بتشتغل_يا_أحمد " becomes " بتشتغل يا أحمد.

3.2. Model

In a complex NLP task like misogyny identification, we need context-aware embedding tools. BERT [3] and derivatives language models provide powerful contextualized embedding.

Recently, some works focused on Arabic language and dialects. We cite AraBERT [5], ARBERT [2], and MARBERT [2]. AraBERT and ARBERT are built on pure MSA datasets, while MARBERT focuses on dialectal Arabic. MARBERT is pre-trained on 1 billion Arabic MSA and DA tweets. In [2] MARBERT showed better performances compared to AraBERT and ARBERT on dialectal Arabic. Thereafter, we choose to use the language model MARBERT to build our classification Model.

After preprocessing the ArMI dataset as described above, we split the dataset into 90% for training and 10% for validation. We have trained MARBERT on the 90% split with a batch size of 32 and a sequence length of 128 for 5 epochs.

4. Results and Discussion

In this section, we present and discuss the results of our experiments on Subtask-1.

4.1. Results on Subtask-1

The evaluation metric used to test our system on Subtask-1 is the accuracy. This metric was specified by the competition organizers. Table 2 lists the results of the classifier built by fine-tuning MARBERT on the training set and tested on the validation set.

4.2. Official Results

Table 3 lists the results performed by our model on the test set as provided by the competition organizers. We observe a large drop in model performance on the test set, which is most likely

Table 2
Results on validation dataset.

	Train loss	Validation accuracy	Validation recall	Validation precision	Validation F1
Results using MARBERT model	0.208	0.876	0.866	0.8762	0.870

Table 3
Results obtained on Subtask-1.

	Accuracy	Precision	Recall	F1
First run results (without cleaning)	0.474	0.5	0.5	0.474
Second run results (with cleaning)	0.483	0.506	0.506	0.483

due to overfitting.

4.3. Error Analysis

We performed extra error analysis for our proposed model results. This analysis aims to find where the model failed to correctly categorize the tweets and tries to discover the causes of this misclassification.
We examine a sample of 50 random misclassified tweets. We discovered various reasons why sarcastic tweets are classified as not misogyny and vice versa. These reasons are summarized as follows:

- **Human annotation** is not perfect since the diversity of annotators' cultures and backgrounds may not be taken into account throughout the annotation process.

- **The absence of context:** In some tweets, the context is missing, making it impossible for our algorithm to grasp the context and accurately forecast the label. Actually, Arabs use some highly offensive words not with the intention to spread hate but by sarcasm.

- **Emojis are not processed:** since we left the emojis in tweets without any kind of preprocessing, we discovered that some emojis had an effect on the categorization process. Actually, 14% of the tweets contain at least one emoji.

5. Conclusion

To identify misogyny on Dialectal Arabic tweets we first proposed a preprocessing strategy. Secondly, we built a classification model based on MARABERT language model which was

selected for the final submission. We observed that considering emojis in the preprocessing step is crucial to the classification performance.

References

[1] 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.

[2] M. Abdul-Mageed, A. Elmadany, E. M. B. Nagoudi, ARBERT & MARBERT: Deep bidirectional transformers for Arabic, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 7088–7105. URL: https://aclanthology.org/2021.acl-long.551.

[3] K. L. Jacob Devlin, Ming-Wei Chang, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, 2019, pp. 4171–4186.

[4] 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.

[5] F. B. Wissam Antoun, H. Hajj, AraBERT: Transformer-based model for Arabic language understanding, in: 2019 Conference of the North American Chapter of thethe 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, Marseille, France, 2020, pp. 9–15.