Detecting Misogyny in Arabic Tweets

Abdusalam Nwesri[1], Stephen Wu[2] and Harmain Harmain[3]

[1]Faculty of Information Technology, University of Tripoli, Tripoli, Libya
[2]School of Biomedical Informatics, UTHealth, Houston, TX USA,
[3]Faculty of Information Technology, University of Tripoli, Tripoli, Libya

Abstract

Systems that can automatically detect offensive content are of great value, for example, to provide protective settings for users or assist social media supervisors with removal of odious language. In this paper, we present three machine learning models developed at University of Tripoli, Libya, for the detection of misogyny in Arabic colloquial tweets. We present the results obtained with these models in the first Arabic Misogyny Identification shared task ArMI'21, a sub track of HASOC@FIRE2021. With our first model (optimized BERT-based pipelines), we placed as the second-ranked team on sub-task A: Misogyny Content Identification, and as the third-ranked team on sub-task B: Misogyny Behavior Identification.

Keywords

Arabic Misogyny detection, hate speech detection

1. Introduction

Public speech that expresses hate or encourages violence toward a person or group based on race, religion, sex, or sexual orientation is defined as hate speech. Expressing feelings of hating women, or believing that men are much better than women is termed as Misogyny.[1] Misogyny is an increasing phenomenon in virtual environments such as social media as people have more freedom to express their feelings with no restrictions than in face-to-face meetings. For example, Facebook reported 31.5 million instances of content with hate speech in the second quarter of 2021.[2] Twitter reported 54 percent increase in the number of accounts violating its hateful conduct policy in a six-month period after July 2019.[3]

Social media companies struggle with the ethical ramifications of misogynistic speech on their platforms. Thus, some companies, such as Twitter, hire a large number of employees to moderate content. However, the huge number of social media posts generated every day makes manual moderation unscalable, so the assistance of automated misogyny detection systems is necessary to enable this type of curation.

Automatically labeling Arabic colloquial tweets as misogynous or non-misogynous is challenging task because the language of tweets is full of syntactic and grammatical flaws, making

Forum for Information Retrieval Evaluation, December 13-17, 2021, India
a.nwesri@uot.edu.ly (A. Nwesri); wu.stephen.t@gmail.com (S. Wu); h.harmain@uot.edu.ly (H. Harmain)

[1]dictionary.cambridge,org
[2]transperancy.fb.com/communit-standards-enforcement/hate-speech/facebook/
[3]blog.twitter.com/en_us/topics/company/2020/new-transparency-center

extraction of text-based features a difficult task. Tweets are short and often consist of few words.

In this paper, we present three models to detect misogyny in Arabic tweets, for the first Arabic Misogyny Identification (ArMI) shared task, a sub track of Hate and Offensive Content Identification (HASOC) at the 2021 Forum for Information Retrieval Evaluation (FIRE@2021).

2. Related Work

Though misogyny detection in Arabic text is a recent topic, some previous work has been done on offensive language and hate speech detection in Arabic.

The first study on abusive language detection in Arabic was done by Ehab A. Abozinadah and Jr. (2015). They tested three machine learning algorithms — Naïve Bayes (NB), Support Vector Machines (SVM), and Decision Tree (J48) classifiers — to detect abusive tweets on a set of 1,300,000 Arabic tweets collected using five swear words. They reported that the NB algorithm was the best performer with an accuracy rate of 90%.

Alakrot et al. (2018) constructed a data set of 167,549 YouTube comments and utilized SVMs to classify comments as either positive or negative. They reported that the SVM classifier achieved 90.05% accuracy.

Husain (2020) tested the impact of the pre-processing phase on the detection of offensive and hate speech for Arabic text. The author used an SVM classifier to identify offensive and hate speech in a data set before and after applying the pre-processing techniques on the original text. The pre-processing techniques improved the classification with an F1 score of 89% for the offensive language detection task and 95% for the hate speech classification task.

Mulki and Ghanem (2021b) built a levantine data set of 6,603 tweets collected from the Twitter accounts of several female journalists who covered the Lebanese protests of October 2019. Tweets in the data set are annotated as misogynous or none. Misogynous tweets are further classified to differentiate between, for example, a threat of violence versus a derailing comment. They used several models to detect misogyny and found that BERT is the best model to classify tweets as misogynous, with an F1 score of 0.88. In the categorical classification they reported that Frenda et al. (2018) model was the best performer with an F1 score of 0.43.

3. Experiment Description

Our experiment was part of the ArMI shared task, a sub track of the HASOC at FIRE 2021. The task aims at identifying misogynistic tweets and recognizing different misogynistic categories in a collection of Arabic (MSA/dialectal) tweets.

3.1. Task Details

ArMI 2021 used a data set composed of (7,866) tweets written in Modern Standard Arabic (MSA) and several Arabic dialects, including Gulf, Egyptian and Levantine (Mulki and Ghanem, 2021c). Participants participated in two sub-tasks. In sub-task A, participants were required to identify a tweet as a misogynistic (`misogyny`) or non-misogynistic (`none`). In sub-task B, participants were required to classify misogynistic tweets to (`discredit`, `derailing`, `dominance`,

`stereotyping & objectification`, `threat of violence`, `sexual harassment`, or `damning`). Two data sets were released, a training set with its gold standard classifications, and a test set with the gold standard withheld. The training set was used to tune detection algorithms and the test set was used to blindly classify new unannotated tweets. More details about tasks are described in (Mulki and Ghanem, 2021a).

4. Experiments

We have participated under the name of University of Tripoli (UoT) with three different runs in each sub-task. Below is the description of these runs in each task:

4.1. Sub-task A (Misogyny Content Identification)

4.1.1. UoT run1: Large BERT-based pipelines

Full pipelines with BERT models (Devlin et al., 2019) at their center were compared for performance on the training set. Words are were segmented using Farasa stemmer.[4] We then then used the American University of Beirut's AraBERT v2 (Antoun et al., 2020) with the BERT-large architecture, pretrained on OSCAR, Arabic Wikipedia, 1.5B words Arabic Corpus, OSIAN Corpus, and Assafir news articles. We then fine-tuned the model on our training data set for misogyny content identification.

4.1.2. UoT run2: Statistical ML Classifiers

We also created a classical machine learning pipeline based on `sklearn` library (Pedregosa et al., 2011). The pipeline consists of 4 stages: a pre-processor, count vectorizer, tf-idf transformer and a multinomialNB classifier. In the text pre-processing stage, any special characters, links, commas, and usernames in @-mentions were removed from the tweets. Hyper parameters were used to tune the transformers and classifier.

4.1.3. UoT run3: Feed-forward networks

We started by removing all non-Arabic characters for the text, then we removed repeated characters leaving only two of the following characters ("و0", "ه", "ا", "ل", "ي", "ف") We then removed the dot character, normalized the different forms of Hamza to a bare Hamza "ا", and split the starting combination of "يا" from any word in the text. This last adjustment was made since this combination is used to address someone in Arabic, is widely used in Arabic hate speech, and is often attached to the following word. We have also normalized wrongly written Arabic phrases widely used in damming someone, such as: "قبح الله", "قبحك الله", "اذن الله", "لعنت الله", "انشاء الله", "ادن الله" . We then normalized "ة" to "ه" and final "ى" to "ى". Finally, we replaced the female addressee pronoun "انتي" with "انت" since most tweets are addressing

[4] https://alt.qcri.org/farasa/

Table 1
Sub-task A results obtained using the training data set

Run	Acc.	Recall	Precision	F1
UoT_run1	0.909	0.903	0.904	0.904
UoT_run2	0.83	0.81	0.82	0.82
UoT_run3	0.841	0.832	0.845	0.838

Table 2
Sub-task B results obtained using the training data set

Run	Acc.	Recall	Precision	F1
UoT_run1	0.769	0.468	0.494	0.474
UoT_run2	0.69	0.55	0.33	0.36
UoT_run2	0.728	0.888	0.508	0.654

females. Using word frequency in the training data set, we have removed a list of 29 tokens chosen based on their frequency in the training data set. The remaining words are transformed to a matrix of numbers based on their tf-idf score in tweets. A 2-layer feedforward neural network was implemented in `keras`. We trained the model with a batch of size 100, and trained the model for 4 epochs. The final F1 score we got is 0.838.

4.2. Sub-task B (Misogyny Behavior Identification)

We treated Sub-task B as a multi-class classification problem, using essentially the same strategy and system for each of the 3 runs, respectively, but training on the more fine-grained descriptions of misogyny behavior: damning, discredit, dominance, sexual harassment, stereotyping & objectification, and threat of violence. Based on the full-pipeline comparisons for UoT_run 1, we selected the highest-performing pipeline, which included basic tokenization and a BERT-large model from Koc University. For run 3, we used the same experimental setup used in sub-task A, however, we used the "binary" mode to transform words to either 0 or 1 based on their presence in the tweets. Results on the training data set are shown in Table 2.

5. Official results

The test set has been released and above runs have been submitted for evaluation to the organizing committee. Table 3 shows results obtained by participating teams including our submitted runs. Our UoT_run1 scored in the 4th position (2nd-best team), while the UoT_run3 and UoT_run2 scored the 12th and the 14th respectively. Our results show that BERT algorithm is the best performer in our runs.

Table 4 shows results of the participants' submitted runs for the sub-task B. Our best performer is again the UoT_run1 at the 8th position (3rd-best team). UoT_run3 and UoT_run2 were in the 11th and the 13th positions respectively.

Table 3
Results of participating teams in sub-task A

Run	Acc.	Recall	Precision	F1
UM6P-NLP_run3	0.919	0.92	0.909	0.914
UM6P-NLP_run2	0.915	0.915	0.905	0.91
UM6P-NLP_run1	0.915	0.911	0.911	0.911
UoT_run1	0.95	0.901	0.899	0.9
SOA_NLP_run1	0.883	0.878	0.876	0.877
BERT	0.88	0.87	0.88	0.87
MUCIC_run1	0.873	0.868	0.864	0.866
SOA_NLP_run2	0.873	0.868	0.865	0.866
(Frenda et. al. 2018)	0.87	0.86	0.86	0.86
MUCIC_run2	0.866	0.86	0.857	0.858
SOA_NLP_run3	0.854	0.846	0.85	0.848
UoT_run3	0.842	0.835	0.831	0.833
iCompass_run1	0.833	0.826	0.82	0.83
UoT_run2	0.827	0.819	0.833	0.822
iCompass_run1	0.508	0.502	0.503	0.499
Isey_run2	0.483	0.506	0.506	0.483
Isey_run1	0.474	0.5	0.5	0.474

Table 4
Results of participating teams in sub-task B

Run	Acc.	Recall	Precision	F1
UM6P-NLP_run2	0.827	0.697	0.647	0.665
UM6P-NLP_run3	0.833	0.717	0.636	0.653
UM6P-NLP_run1	0.816	0.692	0.652	0.651
SOA_NLP_run2	0.764	0.676	0.48	0.531
SOA_NLP_run3	0.745	0.559	0.508	0.526
(Frenda et. al, 2018)	0.77	0.66	0.47	0.52
SOA_NLP_runl	0.78	0.549	0.502	0.519
UoT_run1	0.789	0.541	0.508	0.517
MUCIC_run1	0.765	0.578	0.46	0.497
MUCIC_run2	0.762	0.572	0.456	0.493
UoT_run3	0.73	0.585	0.432	0.468
BERT	0.76	0.54	0.4	0.43
UoT_run2	0.709	0.524	0.382	0.407
iCompassrun2	0.637	0.242	0.248	0.245
iCompass_run1	0.637	0.242	0.248	0.245

6. Conclusion

We have tested three machine learning algorithms in classifying Arabic tweets. AraBert, feedforward networks, and traditional machine learning models have been tested on classifying Arabic tweets as a non-misogynous or misogynous and additionally on classifying misogynic

tweets into six predefined categories. By far the AraBERT algorithm was the best performer with an F1 score of 90% in the first task and 51.7% in the second. In future work, we plan to test the combination of the preprocessing steps we made with the Keras model and the AraBERT approach.

References

Alakrot, A., Murray, L., and Nikolov, N. S. (2018). Towards accurate detection of offensive language in online communication in arabic. *Procedia Computer Science*, 142:315–320. Arabic Computational Linguistics.

Antoun, W., Baly, F., and Hajj, H. (2020). 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*, pages 9–15, Marseille, France. European Language Resource Association.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Ehab A. Abozinadah, A. V. M. and Jr., J. H. J. (2015). Detection of abusive accounts with arabic tweets. *International Journal of Knowledge Engineering*, 1:113–119.

Frenda, S., Ghanem, B., and y Gómez, M. M. (2018). Exploration of misogyny in spanish and english tweets. In *IberEval@SEPLN*.

Husain, F. (2020). Osact4 shared task on offensive language detection: Intensive preprocessing-based approach.

Mulki, H. and Ghanem, B. (2021a). 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.

Mulki, H. and Ghanem, B. (2021b). Let-mi: An arabic levantine twitter dataset for misogynistic language.

Mulki, H. and Ghanem, B. (2021c). Let-Mi: An Arabic Levantine Twitter Dataset for Misogynistic Language. In *Proceedings of the 6th Arabic Natural Language Processing Workshop (WANLP 2021)*.

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