Arabic Misogyny Identification

Fazlourrahman Balouchzahi[1], Grigori Sidorov[1] and
Hosahalli Lakshmaiah Shashirekha[2]

[1]*Instituto Politécnico Nacional, Centro de Investigación en Computación, CDMX, Mexico*
[2]*Department of Computer Science, Mangalore University, Mangalore, India*

Abstract

Social media usually consists of various forms of toxic contents such as Hate Speech (HS) and contents in offensive and abusive languages, in addition to useful and relevant ones. The offensive contents on social media may target a religion, community, individual or group of people, with specific thoughts and beliefs. A category of offensive content targeting women termed as Misogyny is increasing day-by-day and a person/group who shares such content is called a Misogynist. Misogyny detection can be seen as a sub-category of HS and Offensive Language Identification (OLI) tasks in which women and issues regarding them such as their rights are targeted. Despite the several works undertaken for HS and OLI tasks by several researchers, Misogyny detection has been studied rarely even for rich resource languages. To promote Misogyny detection in Arabic language, Arabic Misogyny Identification (ArMI)- a shared task in Forum for Information Retrieval Evaluation (FIRE) 2021 provides the dataset and invites the researches to develop models for Misogyny detection in the given text. The shared task consists of two subtasks which can be modeled as binary and multiclass Text Classification (TC) tasks. This paper describes the models submitted by our team MUCIC to the ArMI shared task. The proposed methodology uses a combination of top frequent char and word n-grams as features to train Machine Learning (ML) classifiers and obtained an accuracy of 0.873 and F1-score of 0.497 for Subtask A and B respectively.

Keywords

Social Media, Hate Speech, Offensive Language, Misogyny Detection, Machine Learning

1. Introduction

The unlimited freedom and anonymity of users on the social media have provided ample of opportunities for several users who wish to share Hate Speech (HS) and abusive content targeting different communities, religions, beliefs, etc. [1, 2]. Knowingly or unknowingly, usually, women, children and the younger generation will be the victims of this hatredness. Women's rights in Middle East countries have always been a concern for the world and feminism. The type of comments on social media that target women and their rights is seen as an action of violence against women and is called as Misogyny. Detecting Misogyny on social media manually is cumbersome and time consuming due to the increased number of users and increase in the Misogyny content. Despite the several works being explored for the automatic detection of HS and OLI in various languages, Misogyny detection has got very less attention even for resource

Forum for Information Retrieval Evaluation, December 13-17, 2021, India
frs_b@yahoo.com (F. Balouchzahi); sidorov@cic.ipn.mx (G. Sidorov); hlsrekha@gmail.com (H. L. Shashirekha)
https://mangaloreuniversity.ac.in/dr-h-l-shashirekha (H. L. Shashirekha)
0000-0003-1937-3475 (F. Balouchzahi)

Table 1
Description of Categories for Subtask B

Category	Description
Damning (Damn)	The tweet contains an offensive invoke or curse against women
Derailing (Der)	The tweet contains texts to validate and justify women abuse and mistreatment
Discredit (Disc)	The tweet contains defamation and offensive language against women
Dominance (Dom)	The tweet contains texts to target equality of men and women rights by implying the superiority of men over women
Sexual Harassment (Harass)	The tweet contains texts describing sexual abuses against women
Stereotyping & Objectification (Obj)	The tweet contains description of women's physical appeal
Threat of Violence (Vio)	The tweet contains a statement of an attention to hostile actions against women
None	The tweet does not contain any misogyny contents

rich languages. Hence, Misogyny detection is not only interesting but challenging also [3]. ArMI[1] [4], a shared task in FIRE 2021[2] is a first step to encourage researchers to develop models for the detection of Misogyny in Arabic texts. With the aim of identifying Misogyny tweets and categorizing them into different Misogynistic behaviors classes, ArMI shared task consists of the following two subtasks:

- **Subtask A - Misogyny Content Identification:** is a binary Text Classification (TC) task where each tweet has to be classified as "Misogynistic (Misogyny)" (if the tweet contains texts against women) or "Non-misogynistic (None)" (otherwise).;
- **Subtask B - Misogyny Behavior Identification:** is a multiclass TC task where each tweet has to be classified into one of the eight categories described in Table 1.

The effectiveness of various types of n-grams as features have been proved by Balouchzahi et al. [1, 5, 6] for Dravidian[3] languages text and code-mixed texts in Dravidian languages for several TC tasks. In continuation with this, to explore the efficiency of n-grams based feature sets for low resource languages, in this paper, we, team MUCIC propose to utilize a combination of 30,000 top frequent char and word n-grams each as feature set to tackle the Misogyny detection challenge in ArMI shared task. The generated feature set transformed into TFIDF vectors is used to train two ML classifiers, namely: Linear Support Vector Machine (LSVM) and Logistic Regression (LR). SVMs are the popular ML classifiers that take advantage of high dimensional feature sets such as n-grams and support various kernel functions. LR is one of the widely employed binary classifier. However, to deal with multiclass TC tasks it utilizes the one-vs-rest (OvR) scheme [1].

The rest of paper is organized as follows: Section 2 gives a summary of the recent literature in

[1] https://sites.google.com/view/armi2021/
[2] http://fire.irsi.res.in/fire/2021/home
[3] https://en.wikipedia.org/wiki/Dravidian_languages

Misogyny detection and Arabic TC tasks followed by the description of Methodology in Section 3. The Experiments and results are mentioned in Section 4 and the paper concludes in Section 5.

2. Related Work

A primary requirement to promote NLP tasks in any language is the availability of annotated datasets. To promote Misogyny detection task in Levantine Arabic language, Mulki et al. [2] collected Tweet-replies to female journalists Tweets during protests that happened in 2019 in Lebanon. The collected Tweets were cleaned to remove non-textual, Arabic-Arabizi mixed Tweets, retweets, duplicate instances, sequence of hashtags and single Tweet. Further, 77,856 Tweets from 7 female journalists' accounts were retrieved using Twitter API[4] and non-Levantine Tweets were removed manually. Services of two female and one male annotator was used to annotate the Tweets into eight categories as mentioned in Table 1 and only 6,603 Tweets were used for annotation. The authors also experimented various ML classifiers as baselines with BOW, LSTM and BERT. BERT outperformed other models and obtained an accuracy of 0.88 and a F1-score of 0.43 for binary and multiclass misogyny detection tasks respectively.

Misogyny detection in the Arabic language has never been studied earlier [4]. However, several HS detection and OLI tasks in Arabic language are experimented and some of them are briefly described below:

Farha et al. [7] explored Deep Learning (DL) and Transfer Learning (TL) approaches for the task of OLI in Arabic language using the SemEval 2020 Arabic OLI shared task dataset. This dataset consists of 7,000 training samples and 1,000 testing samples for two subtasks, namely: Subtask 1 (HS v/s Not-HS) and Subtask 2 (Offensive v/s Not-Offensive). They experimented Bi-directional Long Short Term Memory (BiLSTM) and BiLSTM-Convolutional Neural Network (CNN) as DL models and ULMFiT as TL model. BiLSTM-CNN was used as a multitask learning approach where authors assumed that, if a Tweet contain HS content it is offensive as well. Sentiments labels were also added as an objective in the methodology. Eventually, BiLSTM-CNN obtained best results with F1-scores of 0.904 and 0.737 for OLI and HS detection respectively.

Alshaalan et al. [8] developed an Arabic HS dataset consisting of 9,316 Tweets distributed into five categories, namely: Racist, Religious, Ideological, Tribal, and Regional. Similar to Mulki et al. [2], Twitter API has been used to scrap the Tweets posted during March 2018 to August 2018 based on keywords. The obtained Tweets were pre-processed by converting Emoji to text and removing hashtags, stopwords, white spaces and punctuation followed by filtering spams and normalization and lemmatization of words. They experimented several CNN and Recurrent Neural Networks (RNN) models, BERT transformer as well as ML classifiers using char n-grams for HS detection task in Saudi Twitter sphere and obtained an F1- score of 0.79 for CNN models as the best score among the models.

Another Arabic HS dataset has been developed by Albadi et al. [9] to encourage researchers to work on religious HS detection. The developed dataset covers Tweets related to common religions such as Islam, Christianity, Judaism, and Atheism in Middle East countries. In addition to these religions, the authors also included Sunni and Shia religions. 6,000 tweets (1,000 per religion) were collected using Twitter API and was distributed into six categories, namely: Islam,

[4] http://www.tweepy.org

Sunni, Shia, Christianity, Judaism, and Atheism. Various ML and DL models were experimented as baselines and GRU-based RNN with an F1-score of 0.79 outperformed the other baselines.

3. Methodology

The proposed methodology consists of the following steps:

i) pre-processing the dataset
ii) extracting char and word n-grams from the given text
iii) selecting 30,000 most frequent features in each category and combining them to form a feature set
iv) vectorizing the feature set using TfidfVectorizer[5]
v) training the ML classifiers with the vectors obtained for training set and
vi) evaluating the models using the vectors obtained for the test set

The overview of the proposed methodology is shown in Figure 1.

n-grams are simple and scalable features that are utilized in many NLP tasks. The value of "n" indicates the amount of the context that is captured. Despite consuming less ram and time, n-grams enhance the efficiency of many TC tasks [10]. The range and the total number of features before selecting the frequent ones are presented in Table 2.

The steps to pre-process the dataset are given below:

- **Emoji to text conversion:** all Emojis are converted to corresponding texts in English using de-emojify[6] library. The conversion of Emojis to English words is considered as a better option than removing Emojis as it results in losing important information.
- **Punctuation removal:** since punctuation usually are not informative features for TC, they are removed.
- **Digits removal:** since digits usually are not informative features, they are removed from texts. This also reduces the number of features.
- **Word with small length:** all words of length less than or equal to two are removed to reduce the number of features.
- **Lower casing the words:** lower casing all the uppercase characters is applied only for English words obtained from Emoji to text conversion.

The feature vectors of the train set are used to train LSVM and LR classifiers which are set with default parameters for each subtask of ArMI shared task and the predictions on the test set are submitted to the organizers for evaluation.

4. Experiments and Results

The dataset for ArMI shared task is a collection of Tweets comprised of Gulf, Egyptian and Levantine dialects and Let-Mi [2] dataset is a collection of Levantine dialects Tweets. Rest of the

[5] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html
[6] https://pypi.org/project/demoji/

Table 2
Features' statistics

n-gram type	range	Total No. of features
char	(1, 3)	222,236
word	(2, 5)	153,519

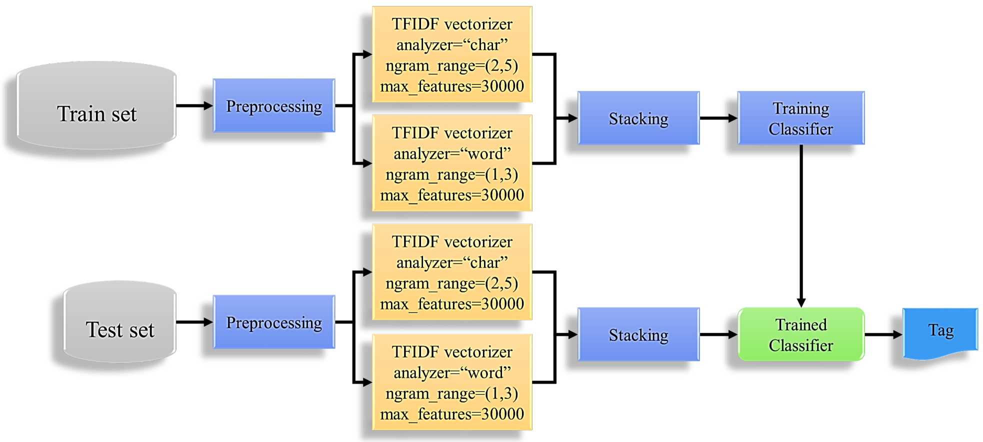

Figure 1: Overview of the proposed method

Table 3
Statistics and label distribution for the training set

Subtask A							
			Misogyny	None			
			4,805	3,061			
Subtask B							
Discredit	Damning	Stereotyping & Objectification	Threat of violence	Dominance	Derailing	Sexual harassment	None
2,868	669	653	230	230	105	61	3,061

multi-dialects Tweets collected from Twitter are based on hashtags, queries and Misogynists' timelines that contain Misogyny content. Participants of the shared task were provided with the training set consisting of 7,866 Tweets (posted during January 2019 - January 2021 and manually annotated by Arabic-native speakers) and test set containing 1,967 tweets (without label) for evaluating the model. The label distribution of the train set for the two subtasks is given in Table 3.

Accuracy and F1-scores are used by the organizers for ranking the models submitted by the participants for Subtask A and B respectively and the results obtained are shown in Table 4. The results illustrates that LR classifier outperformed LSVM with 0.873 accuracy and 0.497 F1-score for Subtask A and B respectively.

The comparison of the accuracies of the models submitted by the participating teams to

Table 4
Performances of the proposed methodology

Subtask	Classifier	Accuracy	Precision	Recall	F1-score
Subtask A	**LR**	**0.873**	**0.868**	**0.864**	**0.866**
	LSVM	0.866	0.860	0.857	0.858
Subtask B	**LR**	**0.765**	**0.578**	**0.460**	**0.497**
	LSVM	0.762	0.572	0.456	0.493

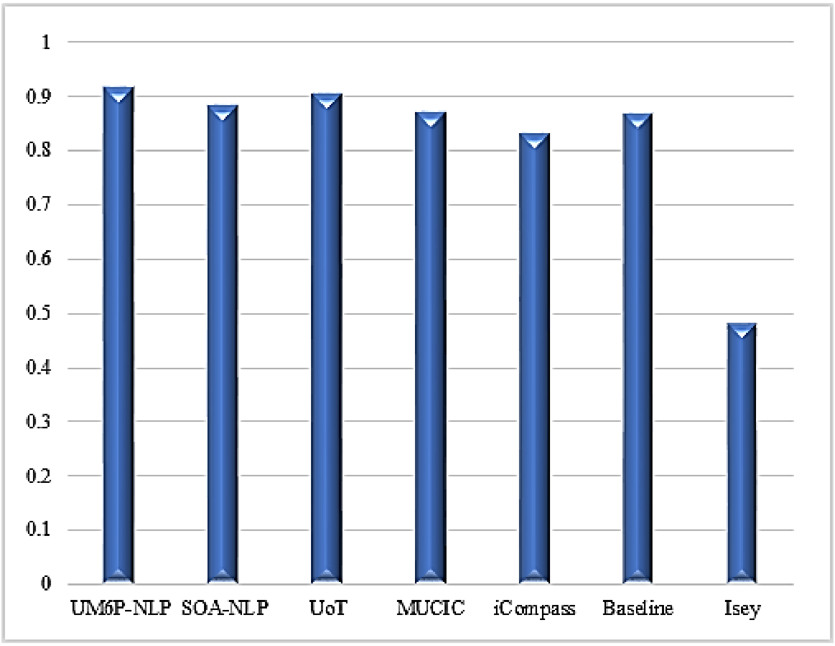

Figure 2: Comparison of accuracies of our model with the top performing teams for Subtask A

Subtask A of the shared task shown in Figure 2 illustrates very competitive results. It can be observed that the difference between the accuracy values of the models (except the models submitted by Isey team) is less than 0.02. The comparison of F1-scores of the models submitted by the participating teams to Subtask B is shown in Figure 3. It can be observed that the difference between the F1-scores of the models (except the model submitted by iCompass) is less than 0.3. The proposed methodology obtained differences of only 0.046 in accuracy and 0.168 in F1-score with the best performing team for Subtask A and B respectively. Analysis of the results also illustrate that all the teams obtained better performance for Subtask A which is a binary TC task.

5. Conclusion and Future Work

This paper describes the model submitted by the team MUCIC to the ArMI shared task which focuses on detecting Misogyny in Arabic language. ArMI shared task consists of two subtasks,

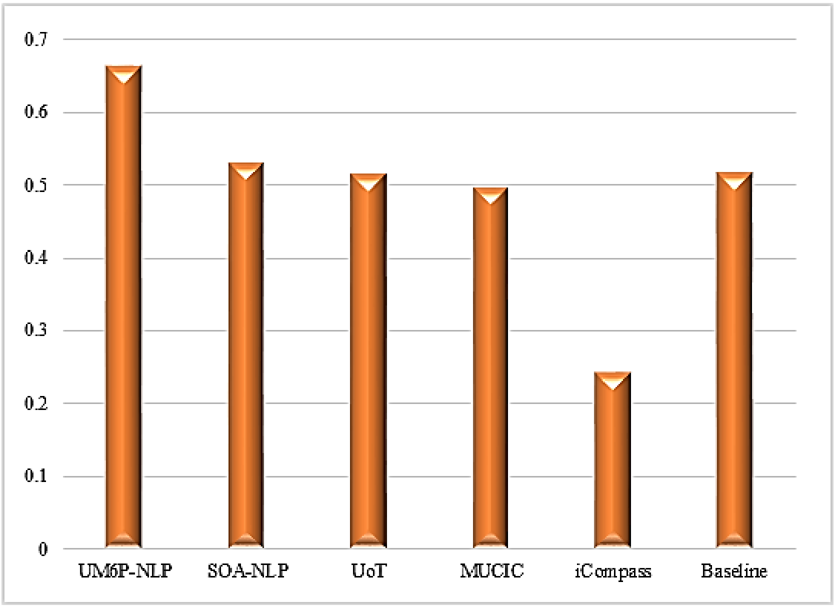

Figure 3: Comparison of F1-scores of our model with the top performing teams for Subtask B

namely: Misogyny Content Identification and Misogyny Behavior Identification which are modeled as binary and multiclass TC tasks respectively. The proposed methodology includes a text pre-processing step followed by generating the most frequent char and word n-grams as features, combining and transforming them to TFIDF vectors. These vectors are used to train two ML classifiers, namely: LSVM and LR. The performances of ML classifiers show very competitive results for the dataset provided by the shared task organizers for both the subtasks. However, LR outperformed LSVM with 0.873 accuracy and 0.497 F1-score in Subtask A and B respectively. Despite the simplicity of the model, our naïve methodology obtained promising results.

The results of our models are expected to be improved further by expanding the experiments on feature engineering part as well as model construction step. Exploring various features, various feature selection algorithms and ensembling various ML classifiers along with exploring TL will be the future work.

Acknowledgments

Team MUCIC sincerely appreciate the organizers for their efforts to conduct this shared task.

References

[1] F. Balouchzahi, B. K. Aparna, H. L. Shashirekha, MUCS@DravidianLangTech-EACL2021: COOLI-Code-Mixing Offensive Language Identification, in: Proceedings of the First

Workshop on Speech and Language Technologies for Dravidian Languages, 2021, pp. 323–329.

[2] H. Mulki, B. Ghanem, Let-Mi: An Arabic Levantine Twitter Dataset for Misogynistic Language, in: Proceedings of the Sixth Arabic Natural Language Processing Workshop, 2021, pp. 154–163.

[3] F. Simona, G. Bilal, M.-y.-G. Manuel, Exploration of Misogyny in Spanish and English tweets, in: Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), volume 2150, Ceur Workshop Proceedings, 2018, pp. 260–267.

[4] 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.

[5] F. Balouchzahi, B. K. Aparna, H. L. Shashirekha, MUCS@ LT-EDI-EACL2021: CoHope-Hope Speech Detection for Equality, Diversity, and Inclusion in Code-Mixed Texts, in: Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion, 2021, pp. 180–187.

[6] F. Balouchzahi, H. L. Shashirekha, MUCS@ Dravidian-CodeMix-FIRE2020: SACO-Sentiments Analysis for CodeMix Text, in: FIRE (Working Notes), 2020, pp. 495–502.

[7] I. A. Farha, W. Magdy, Multitask Learning for Arabic Offensive Language and Hate-Speech Detection, in: Proceedings of the 4th Workshop on Open-source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, 2020, pp. 86–90.

[8] R. Alshaalan, H. Al-Khalifa, Hate Speech Detection in Saudi Twittersphere: A Deep Learning Approach, in: Proceedings of the Fifth Arabic Natural Language Processing Workshop, 2020, pp. 12–23.

[9] N. Albadi, M. Kurdi, S. Mishra, Are they our Brothers? Analysis and Detection of Religious Hate Speech in the Arabic Tswittersphere, in: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2018, pp. 69–76.

[10] F. Balouchzahi, M. D. Anusha, H. L. Shashirekha, MUCS@TechDOfication using FineTuned Vectors and n-grams, in: Proceedings of the 17th International Conference on Natural Language Processing (ICON): TechDOfication 2020 Shared Task, NLP Association of India (NLPAI), Patna, India, 2020, pp. 1–5. URL: https://aclanthology.org/2020.icon-techdofication.1.