ArMI at FIRE 2021: Overview of the First Shared Task on Arabic Misogyny Identification

Hala Mulki[1], Bilal Ghanem[2]

[1] *The ORSAM Center for Middle Eastern Studies, Ankara, Turkey*
[2] *University of Alberta, Edmonton, Canada*

Abstract

This paper provides an overview of the organization, results and main findings of the first shared task on misogyny identification in Arabic tweets. Arabic Misogyny Identification task (ArMI) is introduced within the Hate Speech and Offensive Content detection (HASOC) track at FIRE-2021. The ArMI task combines two related classification subtasks: a main binary classification subtask for detecting the presence of misogynistic language, and a fine-grained multi-class classification subtask for identifying seven misogynistic behaviors found in misogynistic contents. The data provided for this task is a Twitter dataset composed of 9,833 tweets written in modern standard Arabic (MSA) and several Arabic dialects including Levantine, Egyptian and Gulf[1]. ArMI at FIRE-2021 has got a total of 15 submitted runs for Sub-task A and 13 runs for Sub-task B provided by six different teams. The systems introduced by the participants employed various methods including feature-based, neural networks using either classical machine learning techniques, ensemble methods or transformers. The best performing system achieved an F-measure of 91.4% and 66.5% for subtask A and subtask B, respectively. This indicates that misogynistic language detection and misogynistic behaviors identification in Arabic textual contents can be, effectively, addressed using transformer-based approaches.

Keywords

Misogyny identification, Arabic language, Social media

1. Introduction and Motivation

Online misogyny has become a universal phenomenon spread widely across social media platforms. Misogyny is one type of hate speech that disparages a person or a group having the female gender identity; it is typically defined as hatred of or contempt for women [1, 2]. According to [3], based on the misogynistic behavior, misogynistic language can be classified into several categories such as discredit, dominance, derailing, sexual harassment, stereotyping and objectification, and threat of violence. Like their peers all over the world, women in the Arab region are exposed to several forms of online misogyny, through which, gender inequality, violence against women, and underestimation of women are, unfortunately, reinforced and justified [3]. This made the automatic identification

[1] Available at: https://github.com/bilalghanem/armi

FIRE 2021: Forum for Information Retrieval Evaluation, 13th-17th December, 2021
hala.mulki@orsam.org.tr (H. Mulki); bghanem@ualberta.ca (B. Ghanem)
0000-0002-7608-2765 (H. Mulki); 0000-0001-7973-8574 (B. Ghanem)

of online Arabic misogyny very crucial to assist in prohibiting the misogynistic Arabic contents and, thus, enabling Arab women to explore social media safely and express their opinions freely [4]. However, this could not be achieved without the provision of annotated data needed for training and developing automatic misogyny identification systems. While online misogyny detection for Indo-European languages such as English, Spanish and Italian have been addressed by multiple systems presented at several shared tasks [5, 6], Arabic misogynistic language detection has never been tackled in similar tasks because of the lack of Arabic datasets annotated for misogyny. Therefore, the motivation for organizing **Ar**abic **M**isogyny **I**dentification task (ArMI) goes beyond proposing the first shared task of automatic detection of Arabic misogynistic language to enrich the Arabic linguistic resources with a novel type of Arabic toxic content. Thus, We have made the ArMI evaluation dataset publicly available to the research community aiming to develop novel systems able to handle the challenging nature of the Arabic language and Arabic dialects in the context of identifying misogynistic language and misogynistic behaviors. Moreover, through this task, we encourage the participation of several research groups, from both academia and industry, seeking to promote advancement in the state of the art and paving the way in misogyny detection for the Arabic language.

The remainder of this paper is organized as follows: Section 2 presents the shared task description and the sub-tasks included in ArMI. Section 3 provides a detailed review of the evaluation dataset in terms of data collection, annotation, and statistics. In Section 4, the evaluation measures are presented while Section 5 discusses the participating systems, lists and compares their submitted results. Finally, Section 6 concludes the overview study.

2. Task Description

The ArMI shared task aims at identifying the misogynistic language and recognizing different misogynistic behaviors in a collection of Arabic (MSA/dialectal) tweets. The participants can choose to participate in one or both of the following sub-tasks:

- **Sub-task A - Misogyny Content Identification:**
  This sub-task represents a coarse-grained binary classification in which the participating systems are required to classify the tweets into two classes, namely: Misogynistic (Misogyny) and Non-misogynistic (none).

- **Sub-task B - Misogyny Behavior Identification:**
  This sub-task is a fine-grained, multi-class classification of misogynistic behaviors where along with the Non-misogynistic class, the misogynistic tweets from the sub-task A are further classified into seven categories:

  1. Damning: tweets under this class contain cursing content.
  2. Derailing: tweets under this class combine justification of women abuse or mistreatment.

3. Discredit: tweets under this class bear slurs and offensive language against women.
4. Dominance: tweets under this class imply the superiority of men over women.
5. Sexual Harassment: tweets under this class describe sexual advances and sexual nature abuse.
6. Stereotyping & Objectification: tweets under this class promote a fixed image of women or describe women's physical appeal.
7. Threat of Violence: tweets under this class have an intimidating content with threats of physical violence.
8. None: if no misogynistic behaviors exist.

3. ArMI Evaluation Dataset

3.1. Data Collection

The dataset used for evaluation in ArMI shared task contains 9,833 tweets written in Modern Standard Arabic (MSA) and several Arabic dialects including: Gulf, Egyptian and Levantine. The Levantine tweets were derived from Let-Mi dataset [4] which was constructed out of the tweet direct replies posted at the timelines of several Lebanese female journalists who were targeted by online bullying campaigns during the 17 October protests in Lebanon. The multi-dialectal tweets, however, were collected based on anti-women hashtags and specific queries (See Table 1). Moreover, by tracking anti-women hashtags and queries within the "bio" section of Twitter Arab users, we spotted several users who describe themselves as misogynists and scraped the tweets from their public timelines [1]. All the tweets were collected during the period (January 2019 - January 2021) using Twitter API[2].

In order to prepare the collected tweets for annotation, they were normalized by eliminating Twitter-inherited symbols, digits, and URLs. It should be mentioned that as the hashtags encountered within a tweet can indicate a misogynistic content, hashtag symbols were removed while the hashtag words were retained.

3.2. Annotation Process and Evaluation

The dataset used in ArMI shared task is resulted from merging Let-Mi dataset [4] with a collection of Arabic multi-dialectal tweets. Although both datasets were, manually, annotated for misogynistic content and misogynistic behavior using the same annotation guidelines, Let-Mi dataset was annotated by three annotators while two annotators labeled the multi-dialectal dataset. It should be noted that all the annotators are Arabic native speakers and are fluent in the Egyptian, Levantine and Gulf dialects. Based on the definition of misogynistic behaviors in [3], we designed the annotation guidelines such that the eight label categories are identified as follows:

[1] user names were masked throughout the research study.
[2] We used python *Tweepy* library http://www.tweepy.org

Table 1
Anti-women hashtags/queries used for data collection

Hashtags/Queries (Arabic)	Hashtags/Queries (English translation)
#عقليات_البنات	Girls' mentalities
#رقاصات_الاعلام	The belly dancers of media
#إم _نص _لسان	A woman of speech defect
#القصيرات _ولا _الطويلات	Short Women or Tall Women
#الست_نكديه	Women are grumpy
#ريد_بيل	#Red_Pill
هرمونات البنات	Girls' hormones
روحي عالمطبخ	Go to kitchen
تأديب الحريم	Wives' discipline

- Non-Misogynistic (none): tweets are those instances that do not express any hatred, insulting, or verbal abuse towards women.
- Discredit refers to tweets that combine slurring over women with no other larger intention.
- Derailing: used to describe tweets that indicate a justification of women abuse while rejecting male responsibility in an attempt to disrupt the conversation to refocus it.
- Dominance: tweets are those that express male superiority or preserve male control over women.
- Stereotyping & objectification: used to annotate tweets that promote a widely held but fixed and oversimplified image/idea of women. This label also refers to tweet instances that describe women's physical appeal and/or provide comparisons to narrow standards.
- Threat of violence: used to annotate tweets that intimidate women to silence them with an intent to assert power over women through threats of violence physically.
- Sexual harassment: used for tweets that describe actions such as sexual advances, requests for sexual favors, and sexual nature harassment.
- Damning: a misogynistic behavior that is inspired by the Arabic culture. It is used to annotate tweets that contain prayers to hurt women; most of the prayers are death/illness wishes besides praying God to curse women.

Table 2 lists the relevant examples to each class.

Having the data annotated and after the tweets of conflicts were either reduced (three annotators) or agreed upon a unified label for them (two annotators), we ended up with a collection of $9,833$ tweets. A summary of the annotation statistics is presented in Table 3.

Later, we evaluated the judgments of the annotators using inter-annotator agreement measure: Krippendorff's α [7]. According to [4], the calculated Krippendorff's α for was

Table 2
Tweet examples of the annotation labels

Label	Example
Discredit	اخرسي وانضبي وحدة وسخة Shut up! you B**ch
Derailing	من حق الرجل ان يضرب زوجته و هذا هو المهم A man has the right to beat his wife
Dominance	لو منعرفك زلمه منرد عليك We'd have answered you if you were a man
Stereotyping & objectification	انا بكره الستات ف اي حاجة بشوفهم كائنات لا تصلح الا للمطبخ بجد I hate women.. Whatever they do I find them useless.. They can only fit for the kitchen
Threat of violence	ابشري بتصفيتك. عاجل ام أجل يا ديوثه Hey W**re! We will assassinate you sooner or later
Sexual Harassment	خلاص ابقي تعالي اتباسي علشان متبقيش ضيقه الافق Okay.. come and let them kiss you, thus you won't be narrow minded
Damning	أنشالله رصاصة بقلبك May God put a bullet in your heart
None	عرفنا شو كانت كل نواياكي واهدافك من الثورة We have known all your intentions and goals from participating in the revolution

Table 3
Summary of annotation statistics

Annotation Case	#Tweets
Unanimous agreement	8,812
Majority agreement	1,021
Conflicts	53

82.9% which is good. As for the multi-dialectal collection of tweets which were annotated by two annotators, we found that Krippendorff's α is 66.5% which is tentative. Thus, both Krippendorff's α values indicates the consistency of the annotations for each of Let-Mi and the multi-dialectal tweets collection; Consequently, the annotations of ArMI dataset which is constructed out of merging Let-Mi and the multi-dialectal tweets collection and used in the shared task are considered reliable and consistent.

3.3. ArMI Dataset Statistics

Having the annotation process and the evaluation accomplished, we have got a total of 9,833 tweets among which $6,006$ were misogynistic and $3,827$ were non-misogynistic. The

Table 4
Statistics of ArMI Dataset

Class	#Tweets
None	3,827
Discredit	3,586
Damning	836
Stereotyping & objectification	816
Threat of violence	287
Dominance	274
Derailing	131
Sexual harassment	76

Table 5
Misogynistic/non-misogynistic tweets distribution in Train/Test sets

	#Misogynsitic	#Not-Misogynistic	Total
Train	$4,805$	$3,061$	$7,866$
Test	$1,201$	766	$1,967$
Total	$6,006$	$3,827$	$9,833$

Table 6
Misogynistic Behaviors Tweets distribution in Train/Test sets

Class	Train	Test	Total
None	$3,061$	766	$3,827$
Discredit	$2,868$	718	$3,586$
Damning	669	167	836
Stereotyping & objectification	653	163	816
Threat of violence	230	57	287
Dominance	219	55	274
Derailing	105	26	131
Sexual harassment	61	15	76
Total	$7,866$	$1,967$	$9,833$

adopted tweets were distributed unevenly among the misogynistic behavior classes as shown in Table 4.

On the other hand, the distribution of the tweets between training and test sets is given in Table 5. The general class distribution (Misogynistic vs. non-misogynistic) is quite similar, with a proportion of 61% misogynistic tweets in both train and test sets as shown in Table 5. Table 6, however, lists the number of tweets for each misogynistic behavior class in both train and test sets.

4. Evaluation and Metrics

Regarding the submission and systems evaluation process, each participated team is allowed to submit a maximum of three runs. The performance of the submitted approaches for the *Misogynistic Content Identification* task will be evaluated by accuracy. However, the submitted runs of the *Misogynistic Behavior Identification* task will be evaluated using the macro-averaged measures: precision, recall and F1-measure. The final rank of the systems will be sorted by the macro F1-measure.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{1}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{2}$$

$$F1 = \frac{2 * Recall * Precision}{Recall + Precision} \tag{3}$$

$$Accuracy = \frac{True\ Positives}{Total\ number\ of\ instances} \tag{4}$$

5. Participants Methods and Results

Nineteen teams have been registered to the shared task among which **six** submitted their runs. Participants were from 4 different countries: Mexico, Morocco, Tunis, Libya, and India. All team members were from public entities (universities or Research Centers).

Participants used either traditional machine learning approaches or transformers-based models. Some teams preprocessed the input tweets before feeding them to their classifiers by removing URLs, special characters, or duplicate characters within words. Other teams did more than simple text normalization to preprocess the input tweets, where they converted emojis to text to have a textual representation of the tweets. In Table 7, we summarize the participants' approaches in terms of the used models, text representations, preprocessing steps, and post-processing steps.

In the following we summarize the participants approaches:

1. **iCompass** team [8] did not apply any pre or post -process steps. The team used MARBERT pretrained transformer [9], which is an Arabic version of the BERT model trained on Arabic corpora, to identify misogynistic tweets and to recognize the misogynistic behaviors.
2. **IsEy** [10]: Similar to iCompass approach, IsEy team used MARBERT model for both sub-tasks, A and B. In addition to that, the team applied some preprocessing steps to normalize text, such as: removing URLs, user mentions, special characters, duplicate characters, etc.
3. **MUCIC** [11]: Although most of the participated teams used transformers-based approaches, MUCIC team used word and character n-grams models that utilize TF-IDF weighting scheme and select the top 30,000 features from text. After that, the new text representations are fed to a Support Vector Machine (SVM) and

Table 7
A summary of the participants approaches

	iCompass	IsEy	MUCIC	UM6P-NLP	UoT	SOA_NLP
Preprocessing steps						
- Removing URLs		✔			✔	
- Emojis to text			✔			
- Hashtags to words			✔			
- Mentions to words			✔			
- Removing usernames		✔			✔	
- Removing special chars.		✔	✔		✔	
- Removing repeated chars.		✔				
- Processing Arabic letters					✔	
Text representation						
- Bag-of-words/chars			✔		✔	✔
- Lexicons					✔	
Classification Models						
- Logistic Regression			✔			✔
- Naive Bayes					✔	
- Support Vector Machine			✔			✔
- Neural Network					✔	✔
- Arabic BERT						✔
- AraBERT					✔	
- MARBERT	✔	✔		✔		
Postprocessing steps						
- Ensembling				✔		✔

Logistic Regression (LR) models. As a preprocessing step, the authors converted emojis into text, hashtags and mentions decoded into words, and had the special characters removed.

4. **UM6P-NLP** [12] team used MARBERT model as well, but with adding some further task-specific layers on top of it. for the preprocessing, the authors extracted emojis from tweets and added the BERT separator token ([SEP]) between them at encoding time. As a post-processing step, the authors ensembled the predictions from three different versions of MARBERT models.
5. **UoT** [13] team used two main approaches. For the first approach, the text was normalized by removing special characters, URLs, commas, used mentions, and finally, a normalization step was applied to standardize some Arabic letters (Hamza, Yaa', and Taa' Marbota). Later, a word n-grams text representation was used and fed to a Naive Bayes classifier. The authors tested both TF-IDF and word frequency weighting schemes for the text representation. For the second model, the authors used AraBERT model [14] without any auxiliary steps.
6. **SOA_NLP** team [15] used two main models: Arabic-BERT [16] model, and a character n-grams representation with SVM, LR, and Neural Network classifiers (each classifier used in one of the three submitted runs). As a post-process step,

Table 8
Sub-task A participants' results ranked in terms of **accuracy**. Baselines are in italic font.

Team Runs	**Accuracy**	Precision	Recall	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	0.905	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	0.823
UoT_run2	0.827	0.819	0.833	0.822
iCompass_run2	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

the predictions from SVM and LR were ensembled in one of the runs.

In Table 8, we present the results of Sub-task A. We show the performance of each team's runs. We also compare the results of the two tasks with two baselines: AraBERT model [14] and *Frenda et. al, 2018* system [17] which is one of the SOTA systems on the misogyny identification task. It could be noticed that in this task, the system provided by the team *UM6P-NLP* was the best performing system, for all of the three runs. The team used *MARBERT* transformers with an ensembling step. The results show that the top performing systems are using transformers in their best runs. The best word/character n-grams run for this task is *MUCIC_run1*, whose results were better than those obtained by some other transformers-based runs, but still lower than the first baseline.

Regarding sub-task B (see Table 9), the scenario regarding the best performing systems changed slightly; transformer-based models are not performing better than the word/character n-gram models, except for the *UM6P-NLP* team. We can notice that word/character n-gram models ranked better than the other models. This is also noticed with the used baselines where the *BERT* model has a lower performance comparing to the *Frenda et. al, 2018* baseline.

The results obtained for both sub-tasks showed that Arabic misogyny identification SOTA results are in line with English, Spanish [5], and Italian [6] results. Regarding sub-task A, the results of the best performing systems are close (larger than 0.8 of accuracy). On the other hand, the macro-averaged F1-measure results of sub-task B are in between 0.4 to 0.5 for English, Spanish, and Italian, whereas for Arabic it is around 0.67.

Table 9
Sub-task B participants' results ranked in terms of **F1**. Baselines are in italic font.

Team Runs	Accuracy	Precision	Recall	**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_run1	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
iCompass_run2	0.637	0.242	0.248	0.245
iCompass_run1	0.637	0.242	0.248	0.245

6. Conclusion

In this paper, we have presented the results of the first shared task on misogyny identification in Arabic tweets, hosted as a subtrack of HASOC at FIRE-2021. The participants had to identify misogynistic tweets and then to detect the misogynistic behavior within them. Nineteen teams participated in the task and a total of six teams submitted their runs. The systems have been trained on a dataset composed of misogynistic and non-misogynistic tweets for the first sub-task, and for the second sub-task, the tweets have been further annotated with seven different misogynistic behavior classes. ArMI dataset was manually annotated and the inter-annotator agreement was found "good". The methods proposed by the participants ranged from traditional feature-based approaches relying on word or character n-gram features to transformers-based systems. Several transformer models were evaluated, as well as, many classical classifiers were used. Ensemble methods have also been utilized. The best performing system for sub-task A achieved an accuracy value of 0.919, and the best system for sub-task B achieved an F1 score of 0.665. Both systems used transformer-based approaches. Finally, we have made ArMI dataset publicly available to the research community.

References

[1] J. T. Nockleby, Hate Speech, volume 1, Encyclopedia of the American Constitution (2nd ed., edited by Leonard W. Levy, Kenneth L. Karst et al. New York: Macmillan, New York: Macmillan, 2000.

[2] M. E. Moloney, T. P. Love, Assessing online misogyny: Perspectives from sociology and feminist media studies, Sociology compass 12 (2018) e12577.

[3] B. Poland, Haters: Harassment, abuse, and violence online, Lincoln: University of Nebraska Press, 2016.
[4] H. Mulki, B. Ghanem, Let-mi: An Arabic Levantine Twitter dataset for misogynistic language, in: Proceedings of the Sixth Arabic Natural Language Processing Workshop, Association for Computational Linguistics, Kyiv, Ukraine (Virtual), 2021, pp. 154–163.
[5] E. Fersini, P. Rosso, M. Anzovino, Overview of the Task on Automatic Misogyny Identification at IberEval 2018, in: Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), volume 2150, 2018, pp. 214–228.
[6] E. Fersini, D. Nozza, P. Rosso, Overview of the EVALITA 2018 Task on Automatic Misogyny Identification (AMI), in: EVALITA Evaluation of NLP and Speech Tools for Italian, volume 12, 2018, p. 59.
[7] K. Krippendorff, Computing Krippendorff's alpha-reliability (2011).
[8] A. Messaoudi, C. Fourati, M. Kchaou, H. Haddad, iCompass Working Notes for Arabic Misogyny Identification, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.
[9] M. Abdul-Mageed, A. Elmadany, E. M. B. Nagoudi, ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic, arXiv preprint arXiv:2101.01785 (2020).
[10] I. Abbes, E. Nakache, M. Benhajhmida, Context-aware Language Modeling for Arabic Misogyny Identification, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.
[11] F. Balouchzahi, G. Sidorov, H. L. Shashirekha, MUCIC at Arabic Misogyny Identification, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.
[12] A. El Mahdaouy, A. El Mekki, A. Oumar, H. Mousannif, I. Berrada, Deep Multi-Task Models for Misogyny Identification and Categorization on Arabic Social Media, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.
[13] A. Nwesri, S. Wu, H. Harmain, Detecting Misogyny in Arabic Tweets, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.
[14] W. Antoun, F. Baly, H. Hajj, AraBERT: Transformer-based Model for Arabic Language Understanding, in: LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May, 2020, pp. 9–15.
[15] A. Kumar, P. Kumar Roy, J. Prakash Singh, A Deep Learning Approach for Identification of Arabic Misogyny from Tweets, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.
[16] A. Safaya, M. Abdullatif, D. Yuret, KUISAIL at SemEval-2020 Task 12: BERT-CNN for Offensive Speech Identification in Social Media, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 2054–2059.
[17] S. Frenda, B. Ghanem, M. Montes-y Gómez, Exploration of Misogyny in Spanish and English Tweets, in: 3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), volume 2150, CEUR-WS, 2018, pp. 260–267.