Let-Mi: An Arabic Levantine Twitter Dataset for Misogynistic Language

Hala Mulki
The ORSAM Center
for Middle Eastern Studies
Ankara, Turkey
hala.mulki@orsam.org.tr

Bilal Ghanem
University of Alberta
Edmonton, Canada
bilalhgm@gmail.com

Abstract

Online misogyny has become an increasing worry for Arab women who experience gender-based online abuse on a daily basis. Misogyny automatic detection systems can assist in the prohibition of anti-women Arabic toxic content. Developing such systems is hindered by the lack of the Arabic misogyny benchmark datasets. In this paper, we introduce an Arabic Levantine Twitter dataset for Misogynistic language (LeT-Mi) to be the first benchmark dataset for Arabic misogyny. We further provide a detailed review of the dataset creation and annotation phases. The consistency of the annotations for the proposed dataset was emphasized through inter-rater agreement evaluation measures. Moreover, Let-Mi was used as an evaluation dataset through binary/multi-/target classification tasks conducted by several state-of-the-art machine learning systems along with Multi-Task Learning (MTL) configuration. The obtained results indicated that the performances achieved by the used systems are consistent with state-of-the-art results for languages other than Arabic, while employing MTL improved the performance of the misogyny/target classification tasks.

1 Introduction

Social media are the digital tribunes where people can express their thoughts and opinions with the ultimate freedom. However, this freedom comes at a price, especially when it enables spreading abusive language and hate speech against individuals or groups. 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 (Nockleby, 2000; Moloney and Love, 2018). According to (Poland, 2016), 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.

During the last decade, online misogynistic language has been recognized as a universal phenomenon spread across social media platforms such as Facebook and Twitter. Similar to their peers worldwide, women in the Arab region are subjected to several types of online misogyny, through which gender inequality, lower social status, sexual abuse, violence, mistreatment and underestimation are, unfortunately, reinforced and justified. Moreover, in specific contexts, online misogyny may evolve into systematic bullying campaigns, launched on social media to attack and, sometimes, violently threaten women who have a powerful influence over the society; as it is the case with female journalists/reporters (Ferrier and Garud-Patkar, 2018).

The automatic detection of online misogynistic language can facilitate the prohibition of anti-women toxic contents. While many efforts have been spent in this domain for Indo-European languages, the development of Arabic misogyny detection systems is hindered by the lack of the Arabic annotated misogyny resources. Building such resources involves several challenges in terms of data collection and annotation, especially with the drastic lexical variations among the Arabic dialects and the ambiguity introduced by some underrepresented dialects such as the Levantine dialect. Levantine is one of the five major varieties of Arabic dialects. People in Syria, Lebanon, Palestine, and Jordan are considered the native speakers of Levantine with minor differences from one country to another (Sleiman and Menon, 2018).

The 17th October 2019 protests in Lebanon witnessed a heavy use of Twitter especially by the journalists (Aurora, 2019). While covering the protests on the ground, many female journalists were, unfortunately, prone to gender-based abuse receiving misogynistic replies for the tweets they post.

Proceedings of the Sixth Arabic Natural Language Processing Workshop, pages 154–163
Kyiv, Ukraine (Virtual), April 19, 2021.

In this paper, we introduce the first Arabic **Le**vantine **T**witter dataset for **Mi**sogynistic language (LeT-Mi) to be a benchmark dataset for automatic detection of online misogyny written in the Arabic and Levantine dialect[1]. The proposed dataset consists of 6,550 tweets annotated either as neutral (misogynistic-free) or as one of seven misogyny categories: discredit, dominance, cursing/damning, sexual harassment, stereotyping and objectification, derailing, and threat of violence. The credibility and consistency of the annotations was evaluated through using inter-annotator agreement measures: agreement without chance correction(Cohen's Kappa) and overall Inter-annotator agreement (Krippendorff's alpha). In addition, we benchmark SOTA approaches on the dataset to evaluate the effectiveness of current approaches for misogyny detection task in the Arabic language.

2 Related Work

Since no previous Arabic resources were created for misogyny, and given that misogyny language is considered a type of hate speech, we opted to review the non-Arabic misogyny datasets, and the Arabic abusive and hate speech datasets proposed in the literature. We focus on the characteristics of the presented Arabic annotated resources: data source, data size the tackled toxic categories (for non-misogyny datasets), and annotation/collection strategies.

2.1 Misogyny Detection in Non-Arabic Languages

Misogyny detection has been investigated only in a set of languages; precisely in English, Spanish, and Italian. To the best of our knowledge, there are only three datasets in the literature. The work in (Fersini et al., 2018b) proposed the first English and Spanish misogyny detection datasets to organize a shared task on automatic detection of misogyny (AMI). The datasets were collected using three different ways: 1) using seed words like *bi**h*, *w**re*, *c*nt*, etc., to collect tweets; 2) monitoring accounts of potential victims (e.g. well known feminist women); 3) downloading tweets from the history of misogynist Twitter accounts, i.e. accounts that explicitly declared hate against women in their Twitter bio (screen name). After collecting tweets, the authors used CrowdFlower platform [2] to annotate the tweets. The sizes of the final datasets are 3,977 and 4,138 tweets for English and Spanish, respectively. The authors in (Fersini et al., 2018a) organized another shared task on AMI, but with focusing only on English and Italian languages. The authors followed the (Fersini et al., 2018b) approach to build their datasets. The size of the final datasets is 5,000 tweets for each language. Both previous works annotated their datasets for two main sub-tasks: Misogyny Identification (i.e. if the tweet is misogynous or not), and Misogynistic Behaviour and Target Classification (i.e. identifying the type of misogyny and recognition of the targets that can be either a specific user or group of women).

2.2 Arabic Resources for Online Toxicity

According to (Al-Hassan and Al-Dossari, 2019), the toxic contents on social media can be classified into: abusive, obscene, offensive, violent, adult content, terrorism, and religious hate speech. A detailed review of the datasets that covered online toxicity is provided below.

In (Mubarak et al., 2017), two datasets were proposed: a Twitter dataset of 1,100 dialectal tweets and a dataset of 32K inappropriate comments collected from a popular Arabic news site and annotated as obscene, offensive, or clean. The authors in (Alakrota et al., 2018), provided a dataset of 16K Egyptian, Iraqi, and Libyan comments collected from YouTube. The comments were annotated as either offensive, inoffensive, or neutral. The inter-annotator agreement for the whole sample was 71%. The religious hate speech detection was investigated in (Albadi et al., 2018) where a multi-dialectal dataset of 6.6K tweets was introduced. The annotation guidelines included an identification of the religious groups targeted by hate speech. The calculated inter-annotator agreement for differentiating religious hate speech was 81% while this value decreased to 55% when specifying the religious groups targeted by the hate speech. Another type of hate speech was tackled in (Al-Ajlan and Ykhlef, 2018) where the authors presented a Twitter dataset for bullying detection. A dataset of 20K multi-dialectal Arabic tweets was collected and annotated manually with bullying and non-bullying labels. In this study, no annotation evaluation was provided. More recently, a Twitter dataset L-HSAB of 6K tweets was introduced in (Mulki et al., 2019) as a benchmark dataset for automatic detection of

[1] will be made publicly available on Github.

[2] The platform name changed to *Appen*: https://appen.com

Arabic Levantine abusive language and hate speech. The inter-annotator agreement metric denoted by Krippendorff's alpha (α) was 76.5% and indicated consistent annotations.

3 Let-Mi Dataset

Let-Mi can be described as a political dataset as it is composed of tweet replies scraped from the timelines of popular female journalists/reporters during October 17th protests in Lebanon. In the following subsections, we provide a qualitative overview of the proposed dataset, while an annotation evaluation is presented in Section 3.5.

3.1 Data Collection

The proposed dataset was constructed out of Levantine tweets harvested using Twitter API[3]. The collection process relied on scraping tweet replies written at the timelines of several Lebanese female journalists who covered the protests in Lebanon during the period (October 20- November 3, 2019). The accounts of the journalists were selected based on their activity on Twitter and the engagement they get from the people(Aurora, 2019). As a result, we identified seven journalist accounts as resources of the tweet replies, who represent different national news agencies in Lebanon. Initially, we retrieved 77,856 tweets, and then we manually removed the non-Levantine tweets to cope with the paper's goal, which is to provide a Levantine dataset. We also filtered out the non-textual, Arabic-Arabizi mixed tweets, retweets and duplicated instances. In addition, based on regular expressions, we spotted many tweets whose content represents a single hashtag or a sequence of hashtags. We opted to remove these tweets as they were non-informative and were written just to make a hashtag trending. Moreover, to assure that the collected replies are written to target the journalist herself and not a part of side debates among the users within a thread, we removed tweets that mention accounts other than the journalist's. Thus, we ended up with 6,603 direct tweet replies. Table 1 lists the journalist names[4], their news agencies, and the number of tweet replies collected from the timeline of each. In order to prepare the collected tweets for annotation, we normalized them 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, we removed the hashtag symbol while retaining the hashtag words.

Journalist	News Agency	#Tweets
J1	LBC	4,677
J2	Freelancer	1,012
J3	MTV	351
J4	LBC	179
J5	LBC	112
J6	Aljadeed	171
J7	OTV	101

Table 1: Journalists and their tweet replies size.

3.2 Annotation Guidelines

The annotation process for Let-Mi dataset requires labeling the tweets as non-misogynistic, i.e. none or as one of seven misogynistic categories: *discredit, derailing, dominance, stereotyping & objectification, sexual harassment, threat of violence*, and *damning*. It should be noted that, besides the standard misogynistic categories that are usually adopted in the misogyny datasets provided for Indo-European languages (Fersini et al., 2018b,a; Anzovino et al., 2018), we added the "*damning*" category which represents a novel misogynistic behavior inspired by the Arabic culture. It can be described as the Arabic version of cursing where people ask God to make a woman ill, handicapped, dead, hurt, etc. Given the difficulty associated with the recognition of misogynistic behaviors (Schmidt and Wiegand, 2017), they were clearly defined through our designed annotation guidelines. This enabled the annotators to have a unified perspective about misogynistic language categories and contributed to improving the inter-annotator agreement scores. Based on the definition of misogynistic behaviors in (Poland, 2016), we designed the annotation guidelines for Let-Mi dataset such that the eight label categories are identified as follows:

- 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 re-

[3] We used python *Tweepy* library http://www.tweepy.org.

[4] We masked the journalists' names referring to them as J1, J2,... etc.

Label	Example
None	لما بكون الأمر واضح وبتاخدي لمكان آخر بيضرب مصداقيتك When a matter is clear, and you manipulate it, it affects your credibility
Discredit	انت خلقتي حمارة وعشتي جحشة ولح تموتي بغلة You were born as a donkey, and you're living as a colt, and you will die as a mule
Derailing	هيدي معاقه وصار لازم العالم تتمسخر عليها بلكي بتحس عا حاله عميلة الاعلام She is retarded, people must mock her, maybe she realizes what she's done...Media Rat!
Dominance	لو منعرفك زلمه منرد عليك We'd have answered you if you were a man
Stereotyping & Objectification	مش حضرتك لبتقرري هيدا شي انتي فيكي تقرري شو تلبسي وكيف تتمكيجي مش اكتر You're not the one who can decide such a thing, you can just decide what to wear and how to put your makeup
Threat of violence	والله والله العظيم وين بلاقيك بدي اقتلك قواص ذبح حيالله I swear by God that wherever I find you I will shoot you or slay you
Sexual Harassment	و سوف نغتصبك أيضا و معك J1 الواطية We will rape you and rape degenerate J1 as well
Damning	الله يلعنك ويحرقك ويبعتلك سرطان براسك May God curse and burn you and put cancer in your brain

Table 2: Tweet examples of the annotation labels.

jecting male responsibility with an attempt to disrupt the conversation in order 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: 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.

In addition, we provide target annotation for each tweet found to be misogynistic (fall in one of the seven misogynistic categories), therefore we asked the annotators to tag each misogynistic tweet as belonging to one of the following two target categories:

- Active (individual): the text includes offensive tweets purposely sent to a specific target (explicit indication of addressee or mention of the journalist name);
- Passive (generic): it refers to tweets posted to many potential receivers (e.g. groups of women).

3.3 Annotation Process

The annotation task was assigned to three annotators, one male and two females Levantine native speakers. Besides the previous annotation guidelines, and based on the domain and context of the proposed dataset,we made the annotators aware of specific phrases and terms which look normal/neutral while they indicate toxicity. These phrases/terms are either related to the Lebanese culture or developed during the protests in accordance with the incidents. For example, "بلوطة" (*oak*), which represents a tree type is actually derived from a Lebanese idiom and usually used to describe someone as a liar. Also, the word "سحسوح" (*a slap*) has recently emerged and used during the protest incidents to express an act of violence.

Annotation Case	#Tweets
Unanimous agreement	5,529
Majority agreement (2 out of 3)	1,021
Conflicts	53

Table 3: Summary of annotation statistics.

Having all the annotation rules setup, we asked the three annotators to label the 6,603 tweets as either non-misogynistic or one of the seven misogynistic language categories. When exploring the annotations obtained for the whole dataset, we faced three cases:

1. Unanimous agreement: the three annotators annotated a tweet with the same label. This was encountered in 5,529 tweets.

2. Majority agreement: two out of three annotators agreed on a label of a tweet. This was encountered in 1,021 tweets.

3. Conflicts: each annotator annotated a tweet differently. This case was found in 53 tweets.

After excluding the tweets having three conflicted judgments, the final version of Let-Mi composed of 6,550 tweets. A summary of the annotation statistics is presented in Table 3.

3.4 Annotation Results

With the annotation process accomplished, and considering the annotation cases in Table 3, the final label of each tweet was determined. For tweets falling under the unanimous annotation case, the final labels were directly deduced, while for those falling under the majority annotation case, we selected the label that has been agreed upon by two annotators out of three. Consequently, we got 3,388 non-misogynistic tweets and 3,162 misogynistic tweets where the latter were distributed among the seven misogynistic categories. A detailed review of the statistics of Let-Mi final version is provided in Table 4 where Voc. denotes the vocabulary size for each class.

Label	#Tweets	#Words	Voc.
None	3,388	28,610	12,763
Discredit	2,327	16,587	6,817
Stereotyping & objectification	290	2,235	1,426
Damning	256	1,479	868
Threat of violence	175	1,356	971
Derailing	59	497	391
Dominance	38	292	228
Sexual harassment	17	94	87

Table 4: Tweets distribution across Let-Mi classes.

To identify the words commonly used within misogynistic contexts, we investigated the lexical distribution of the dataset words across the misogynistic classes. Therefore, we subjected Let-Mi to further normalization, where we removed stopwords based on a manually built Levantine stopwords list. Later, we identified the five most frequent words and their frequencies in the misogynistic classes as seen in Table 5, where Dist. denotes the word's distribution in a specific class.

Dominance	Dist.	Violence	Dist.
السيد	3.77%	J1	1.90%
Al-Sayyid		*J-name*	
راسك	3.42%	راسك	1.17%
your head		*your head*	
اشرف	3.41%	موتي	0.66%
more honest)		*be dead*	
صرمايتو	1.37%	الواطية	0.51%
his slippers		*degenerate*	
داعس	1.03%	سحسوح	0.44%
step on		*slap*	

Table 5: Distribution of top 5 frequent terms in dominance/threat of violence classes.

3.5 Annotation Evaluation

We evaluated the annotations using two inter-annotator agreement measures: Cohen's kappa (McHugh, 2012) and Krippendorff's α (Krippendorff, 2011). For the obtained annotations, we found that pairwise Cohen's Kappa between the two female annotators (F1,F2) was 86.2%, while for the annotator pairs (male, female), the value decreased to 81.6% and 80.8%, respectively. Moreover, the calculated Krippendorff's α was 82.9% which is considered "good" (Krippendorff, 2011) and indicates the consistency of the annotations. Aiming to investigate the disagreement among the annotators, we explored the tweets for which the annotators gave different judgments. A sample of the tweets along with the annotations assigned by the annotators is listed in Table 6 where F and M denote female and male, respectively.

As seen in Table 6, the disagreements spotted among the annotators can be justified by the gender of the annotator where both female annotators had the same judgment for tweets that describe women-related issues (e.g. Botox) besides they were more sensitive to violent threats compared to the male annotator. On the other hand, when it comes to sarcastic tweets, regardless of the gender, the annotators provided different class labels due to the normal difference of sense of humor among them.

Tweet	F1	F2	M1
يلا انزلي انتي جاي عبالك سحسوح مرتب شطة *Come on! Try to be at the protests, it seems that you'd like a perfect slap*	violence	violence	derailing
عقبال القتلة الجاية هههه *Wishing you a next beating hhhh*	violence	derailing	violence
تروح تنقبر وتلتهي بالبوتوكس وموتوكس وتترك السياسة يحرق روحا *She should be taking care of her Botox and abandon politics; Damn her soul!*	stereotyping	stereotyping	damning

Table 6: Disagreements among annotators.

4 Experiments and Evaluation

In this section, we describe our experiments on the Let-Mi data. We evaluate the performance of SOTA models on our dataset. We design our experiments at three levels (tasks):

1. Misogyny identification (Binary): tweets contents are classified into *misogynistic* and *non-misogynistic*. This requires merging the seven categories of misogyny into the misogyny class.

2. Categories classification (Multi-class): tweets are classified into categories: *discredit, dominance, damning, derailing, sexual harassment, stereotyping and objectification, and threat of Violence*, or *non-misogynistic*.

3. Target classification (Multi-class): tweets are classified into either *passive, active*, or *non-misogynistic*.

4.1 Models

To present a diverse evaluation on the Let-Mi dataset, we test various approaches that use different text representations. In the following, we present the used approaches:

1. BOW + TF-IDF: word ngrams model with TF-IDF weighting scheme. We test several classifiers on a validation part, and we select Naive Bayes classifier[5].

2. Frenda et al. (2018) model: we use one of the SOTA systems on misogyny identification task. This model combines character ngrams with several lexicons created by the authors to highlight important cues in the misogynistic tweets. For instance, the *Stereotypes* lexicon contains words related to the stereotypes about women, like cooking, taking care of children, etc. Since this approach was proposed for English and Spanish languages, we used Google Translation API to translate the lexicons to the Arabic language. It is worthy of mentioning that we discard the *Abbreviations* lexicon as it is untranslatable. Following the authors' configurations, and we use an ensemble technique (majority voting) to combine the predictions from three different classifiers (NB, SVM, and LR).

3. LSTM: A Long Short-Term Memory (LSTM) neural network that uses Aravec Arabic word embeddings (Soliman et al., 2017) to represent texts. The output representation of the LSTM is fed to a softmax layer.

4. BERT: is a text representation model that showed leading performance on multiple NLP benchmarks (Devlin et al., 2019). Since BERT was trained for the English language, we used AraBert (Antoun et al., 2020) which is a version trained on Arabic texts. We fed the hidden representation of the special [CLS] token that BERT uses to summarize the full input sentence, to a softmax layer.

4.2 Experimental Setup

Given that Let-Mi dataset is not balanced, especially in the category classification task, we split the dataset into training, validation, and test sets using stratified sampling technique; we take a random 20% of the tweets from each class for the validation and test sets. Discarding the classes' size in the splitting process may affect the minority classes (e.g., sexual harassment). For the preprocessing of the text, we remove all special characters, URLs, users mentions, and hashtags to ensure that the evaluation models are not biased to any Twitter-inherited symbol. Regarding the experiments' metrics, we used accuracy and macro precision, recall, and F1 score.

[5] We tested Logistic Regression (LR), Random Forest (RF), Naive Bayes (NB), and Support Vector Machine (SVM) classifiers

4.3 Results

In Table 7, we present the results of the misogyny identification task. Considering the F1 metric, the best result of this task is achieved using the BERT model, performing better than the rest of the models. We can see that (Frenda et al., 2018) model performs slightly better than the BOW and the LSTM models. The results show that the LSTM-based model has the lowest performance.

Model	Acc.	Prec.	Rec.	F1
Majority class	0.52	0.26	0.50	0.34
BOW + TF-IDF	0.84	0.84	0.84	0.84
Frenda et al. (2018) model	0.85	0.86	0.85	0.85
LSTM	0.82	0.82	0.82	0.82
BERT	0.88	0.89	0.88	**0.88**

Table 7: Results of the misogyny identification task.

Regarding the category classification task, the results in Table 8 show a different scenario. The best performing model in misogyny identification task, BERT, performs weaker than BOW and (Frenda et al., 2018) models. We speculate that this is due to the number of instances for each class. Recent studies (Edwards et al., 2020) showed that the performance of the pre-trained models (e.g. BERT) decreases marginally when the number of instances in the training data is less than ∼5K, and this was evident for multi-class classification datasets. In this task, the total size of the training data is ∼4K, and for most of the classes, the maximum number of instances is ∼200.

Model	Acc.	Prec.	Rec.	F1
Majority class	0.52	0.06	0.12	0.09
BOW + TF-IDF	0.79	0.52	0.37	0.41
Frenda et al. (2018) model	0.81	0.62	0.38	**0.43**
LSTM	0.75	0.42	0.3	0.33
BERT	0.81	0.44	0.33	0.35

Table 8: Results of the categories classification task.

For the third task, the results in Table 9 show that all the models have a very competitive performance, with a small improvement for the (Frenda et al., 2018) model.

Model	Acc.	Prec.	Rec.	F1
Majority class	0.52	0.17	0.33	0.23
BOW + TF-IDF	0.82	0.86	0.77	0.81
Frenda et al. (2018) model	0.85	0.88	0.8	**0.83**
LSTM	0.82	0.82	0.81	0.82
BERT	0.83	0.85	0.82	0.82

Table 9: Results of the target classification task.

4.4 Multi-task Learning

In multi-task learning (MTL), a set of relevant tasks (two or more) are involved in the training process of a model to improve the performance on each of them (Caruana, 1997). MTL enables a model to use cues from various tasks to improve each of them performance, or of a target task (Zhang and Abdul-Mageed, 2019). MTL has been employed in several NLP tasks (Kochkina et al., 2018; Majumder et al., 2019; Samghabadi et al., 2020). In this work, we use our three tasks in an MTL configuration to investigate whether the performance of the model on each of them will improve. Therefore, we use the BERT model in our experiment. In Figure 1 we illustrate the MTL configuration with an example.

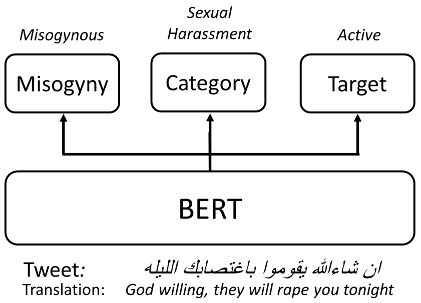

Figure 1: BERT model in an MTL configuration.

The MTL experiment results listed in Table 10 show that MTL improves the performance of both misogyny and target classification tasks clearly, with an average improvement of %4 on the F1 metric. However, the category classification task's performance does not improve and decreases slightly compared to the BERT performance without MTL (1% drop).

Task	Acc.	Prec.	Rec.	F1
Misogyny identification	0.9	0.9	0.9	0.9
Categories classification	0.82	0.42	0.32	0.34
Target classification	0.89	0.89	0.88	0.88

Table 10: Results of the MTL experiment using BERT model.

4.5 Error Analysis

Given the previous sections' results, we can notice that the models' performances in the second task are weak compared to the other two tasks. Thus, in this section, we investigate the causes of the miss-classified cases in the second task. A manual analysis allows us to see which category is the most difficult to be dealt with in the best performing system. Figure 2 reports the confusion matrix of gold labels (y-axis) vs. predicted labels (x-axis).

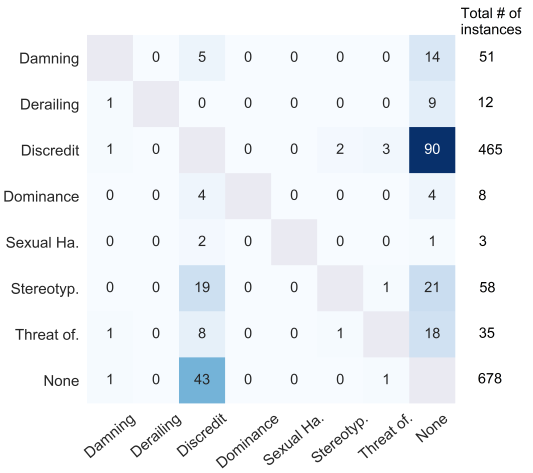

Figure 2: Confusion matrix of errors.

The manual analysis of the tweets shows that classification errors are due to five main factors:

1. **Weak predictions**: although the tweets have been annotated correctly, the model cannot always detect the misogyny category. The cause of this type of errors could be due to the model's generalizability; it could be noted that the model is not able to learn few training cases of some misogynous textual patterns; Besides, this type of errors is evident in the categories having the lowest number of instances in the training data such as *derailing*, *dominance*, and *sexual Harassment*.

2. **Mixed labels**: few tweets can belong to more than one misogyny category, but the annotators chose the most dominant one. In this type of errors, we find that the model sometimes predicts the other possible category. For instance, the tweet: "**انتي مذيعه يا تقبريني خليكي بتلبي عزايم وولائم**"[6] annotated as *Stereotyping & Objectification* because the author is offending the target women by asking her to stay at home to cook food. It can also be considered discredit since the tweet's beginning is discrediting the targeted woman for her performance as a journalist.

3. **Sarcasm**: many tweets annotated under one of the misogyny categories due to their sarcastic meaning. Unfortunately, the model is not able to detect the sarcastic sense. For instance, the tweet: "**كيف صرتي مذيعة؟**"[7] is sarcastically offending the target woman by meaning that she should not be a journalist.

4. **Unique damnation phrases**: some tweets use a unique way or a phrase to curse and attack women. Similar to other languages, the Arabic language contains many phrases for damnation. Apart from the well known phrases, some authors created their own phrases that need some cultural knowledge to be understood. An example on these cases is the tweet: "**ان شاءالله بسقطو زلاعيمك على المسالك البولية**"[8]. This tweet's indirect meaning is that the author prays to God that the journalist will be muted person.

5. **False misogyny cases**: as we are interested in detecting misogynistic tweets, Let-Mi also contains none-misogynistic tweets to classify both groups. Many of the none-misogynistic tweets are neutral tweets (e.g. opinions, questions, news, etc.), but there are also general offensive tweets; not against women. Some of these tweets are offences to the misogynist authors. The analysis shows that the model sometimes cannot discriminate between offensive tweets that target women and the universal ones. An example on this type is the following tweet: "**J1 اشرف منكو ومن اللي خلفكو انتو اللي واطيين يا عبدت الفاسدين والحرامية**"[9].

5 Conclusion and Future Work

This work proposes Let-Mi, the first misogyny detection dataset for the Levantine dialects of the Arabic language. The dataset is annotated manually by a set of Levantine native speakers. We present a detailed description of the whole annotation process. Additionally, we present an experimental evaluation of several machine learning systems, including SOTA systems. Also, we employ an MTL configuration to investigate its effect on the tasks. The

[6] Explanatory translation: are you a journalist! stay at home to cook food for others.

[7] Translation: How you became a journalist?

[8] Literal translation: God willing, your pharynx will fall on the urinary tract.

[9] Literal translation: J1 is more honest than you and your parents, you who are dirty, and slaves of the corrupt and thieves.

results show that the performances of the used systems are consistent with SOTA results on the other languages, and MTL improved the performance of the used model on two of the proposed tasks. In future work, we plan to create a multi-label version of our dataset for the categories classification task as we found that some tweets can be considered under two or more categories. Moreover, since we found many misogynistic tweets that are sarcastic in our analysis, we plan to study the correlation between sarcasm and misogyny tasks.

References

Monirah A. Al-Ajlan and Mourad Ykhlef. 2018. Optimized twitter cyberbullying detection based on deep learning. In *Proceedings of the 21st Saudi Computer Society National Computer Conference (NCC)*, pages 52–56.

Areej Al-Hassan and Hmood Al-Dossari. 2019. Detection of hate speech in social networks: A survey on multilingual corpus. *Computer Science & Information Technology (CS & IT)*, 9(2):83–100.

Azalden Alakrota, Liam Murray, and Nikola S.Nikolov. 2018. Dataset construction for the detection of antisocial behaviour in online communication in arabic. *Procedia Computer Science*, 142:174–181.

Nuha Albadi, Maram Kurdi, and Shivakant Mishra. 2018. Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. In *Proceedings of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 69–76.

Wissam Antoun, Fady Baly, and Hazem Hajj. 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.

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *International Conference on Applications of Natural Language to Information Systems*, pages 57–64. Springer.

Data Aurora. 2019. Lebanon protest statistics. "https://lebanonprotests.com/report". Online; accessed 10-January-2021.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186.

Aleksandra Edwards, Jose Camacho-Collados, Hélène De Ribaupierre, and Alun Preece. 2020. Go Simple and Pre-Train on Domain-Specific Corpora: On the Role of Training Data for Text Classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5522–5529.

Michelle Ferrier and Nisha Garud-Patkar. 2018. Trollbusters: Fighting online harassment of women journalists. In *Mediating Misogyny*, pages 311–332. Springer.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018a. Overview of the EVALITA 2018 Task on Automatic Misogyny Identification (AMI). In *EVALITA Evaluation of NLP and Speech Tools for Italian*, volume 12, page 59.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018b. 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, pages 214–228.

Simona Frenda, Bilal Ghanem, and Manuel Montes-y Gómez. 2018. Exploration of Misogyny in Spanish and English Tweets. In *3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*, volume 2150, pages 260–267. CEUR-WS.

Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3402–3413.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Navonil Majumder, Soujanya Poria, Haiyun Peng, Niyati Chhaya, Erik Cambria, and Alexander Gelbukh. 2019. Sentiment and sarcasm classification with multitask learning. *IEEE Intelligent Systems*, 34(3):38–43.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, 22(3):276–282.

Mairead Eastin Moloney and Tony P Love. 2018. Assessing online misogyny: Perspectives from sociology and feminist media studies. *Sociology compass*, 12(5):e12577.

Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive language detection on arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56.

Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. 2019. L-hsab: a levantine twitter dataset for hate speech and abusive language. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 111–118.

John T. Nockleby. 2000. *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.

Bailey Poland. 2016. *Haters: Harassment, abuse, and violence online*. Lincoln: University of Nebraska Press.

Niloofar Safi Samghabadi, Parth Patwa, PYKL Srinivas, Prerana Mukherjee, Amitava Das, and Thamar Solorio. 2020. Aggression and misogyny detection using bert: A multi-task approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 126–131.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10.

Jamillah Sleiman and Mythili Menon. 2018. The changing of arabic terminology in times of war and displacement. *Proceedings of the Twenty-Ninth Western Conference on Linguistics*.

Abu Bakr Soliman, Kareem Eissa, and Samhaa R. El-Beltagy. 2017. Aravec: A Set of Arabic Word Embedding Models for Use in Arabic NLP. *Procedia Computer Science*, 117:256 – 265. Arabic Computational Linguistics.

Chiyu Zhang and Muhammad Abdul-Mageed. 2019. Multi-task Bidirectional Transformer Representations for Irony Detection. *arXiv preprint arXiv:1909.03526*.