Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2282–2291
Marseille, 20-25 June 2022

ArMIS - The Arabic Misogyny and Sexism Corpus with Annotator Subjective Disagreements

Dina Almanea and Massimo Poesio
Queen Mary University of London
{d.almanea, m.poesio}@qmul.ac.uk

Abstract

The use of misogynistic and sexist language has increased in recent years in social media, and is increasing in the Arabic world in reaction to reforms attempting to remove restrictions on women lives. However, there are few benchmarks for Arabic misogyny and sexism detection, and in those the annotations are in aggregated form even though misogyny and sexism judgments are found to be highly subjective. In this paper we introduce an Arabic misogyny and sexism dataset (ArMIS) characterized by providing annotations from annotators with different degree of religious beliefs, and provide evidence that such differences do result in disagreements. To the best of our knowledge, this is the first dataset to study in detail the effect of beliefs on misogyny and sexism annotation. We also discuss proof-of-concept experiments showing that a dataset in which disagreements have not been reconciled can be used to train state-of-the-art models for misogyny and sexism detection; and consider different ways in which such models could be evaluated.

Keywords: misogyny, sexism, subjectivity, disagreement

1. Introduction

The growth in recent years of social networks and online discussions groups has unfortunately been exploited by some individuals to spread offensive views about people with certain characteristics (religion, ethnicity, women, etc). Although the offenders are not a majority, these offensive views can affect and stigmatize individuals such as feminists attacked in misogynous and sexist posts. This phenomenon has motivated much research on detecting offensive and hate texts online.

Studies have found that women are especially subject to offensive language on social media (Frenda et al., 2019). (According to Amnesty International, an offending remark towards a woman politician or journalist appears on social media every 30 seconds.[1]) Our work is concerned with detecting misogynistic and sexist tweets, focusing in particular on Arabic. The spread of this type of offensive language has recently increased in Saudi Arabia due to the changes brought about by the Saudi government's Vision 2030 project aiming to relax restrictions on women activities[2], for example laws allowing women to drive and/or to be employed in high-level government positions. These changes have originated a major debate in Saudi Arabia: some people agree with these reforms, but others do not. Unfortunately such debates often involve the use of offensive language or even hate speech.

The research in this paper is motivated by a feature of work on misogynistic and sexist language detection that has not attracted much attention so far–namely, the fact that judgments about what counts as offensive appears to very much depend on certain characteristics of those making the judgments. Women appear to find different texts misogynistic than men, and conservatives from liberals. This suggests that such differences should be taken into account when annotating misogynistic and sexist text. In our data collection we also found many examples suggesting that misogynistic and sexist language takes a different form in Arabic than in Western languages, due to the different impact of cultural and religious beliefs. In this study we aim to investigate how misogynistic and sexist judgments in Arabic text are affected by two characteristics of annotators: gender and religious beliefs (whether the coder is religiously liberal, moderate or conservative). We introduce ArMIS, a novel Arabic misogyny and sexism dataset that was annotated by annotators with different degree of Islamic beliefs. We report on a series of annotation experiments testing our preliminary hypotheses about the effect of gender and different religious beliefs on the annotation; our results show a significant effect of beliefs on disagreement between annotators. Finally, we discuss different approaches for training and evaluating misogyny and sexism detection models on data in which all individual judgments are preserved.[3]

2. Background

In this Section we briefly summarize first the literature on offensive and hate speech in general focusing in particular on Arabic resources and on language resource creation. We then discuss the literature more specifically on misogyny/sexism, and finally the literature on

[1] https://www.amnesty.org.uk/press-releases/women-abused-twitter-every-30-seconds-new-study

[2] https://english.alarabiya.net/features/2020/10/18/Top-10-moments-for-Saudi-Arabian-women-since-Vision-2030

[3] Note: due to the nature of this paper, the examples include offensive language which doesn't reflect the authors' views.

disagreements due to bias arising in this area and on developing models from datasets containing bias.

2.1. Hate speech and offensive language

Much recent research has focused on detecting offensive speech and hateful expressions. 'Abusive language' is a general term covering, e.g., rudeness, disrespect and insults, which is difficult to define precisely. This is also true for hate speech in particular (Poletto et al., 2021), although the definition by Nockleby (2000)–"any communication that disparages a person or a group on the basis of some characteristic such as race, colour, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic"–is widely used. The overlap between abusive language and related phenomena is discussed by Waseem et al. (2017). One of the earliest hate speech datasets for English (16k tweets related to sexism and racism) was created by Waseem and Hovy (2016). For Arabic, Mubarak et al. (2017) created two datasets–one of tweets in dialectal Arabic, the other of Aljazeera news–labeled as "obscene", "offensive", or "clean". Other datasets were created by Albadi et al. (2018) (religious hate language in tweets) and Mulki et al. (2019) (the L-HSAB dataset of tweets in Levantine Arabic). The Arabic offensive dataset presented by Mubarak et al. (2020) was used in the OffensEval-2020 shared task on Multilingual Offensive Language Identification (Zampieri et al., 2020), together with datasets in other languages.

2.2. Misogyny and Sexism

Women are frequently the target of hate speech (Poland, 2016; Davidson et al., 2017) and more in general of **misogynistic language** (Bartlett et al., 2014; Hewitt et al., 2016), a category which overlap with sexism towards women (Frenda et al., 2019) [4].

Misogyny has been defined as "hate or prejudice against women, which can be linguistically manifested in numerous ways, ranging from less aggressive behaviours like social exclusion and discrimination to more dangerous expression related to threats or violence and sexual objectification" (Anzovino et al., 2018). As argued by Poletto et al. (2021), although hate speech is a subset of abusive language, not all misogynistic text aims to promote violence or threats towards women, which would put it in the hate speech class according to their definition. Thus, misogynous text can be thought of as a type of abusive language, and can include hate speech, aggressive, and offensive language. A taxonomy of misogynistic language behaviors in social media was proposed by Anzovino et al. (2018) based on the work by Poland (2016) which includes: (i) discrediting women without any further intentions, (ii) stereotyping and objectifying women to subordinate them or to characterize negatively their physical appearance, (iii) sexually harassing them and threatening them with violent intentions, (iv) dominating women and preserving male control over them and (v) justifying abuse over women and disrupting discussions on those responsible for it.

Misogyny detection has attracted a lot of attention in recent years and a number of shared tasks have been organized. These include the shared tasks on Automatic Misogyny Identification (AMI) in different languages at Evalita 2018 covering English and Italian (Fersini et al., 2018a) and at IberEval 2018 covered English and Spanish (Fersini et al., 2018b). The AMI-2018 tasks involved two hierarchically organized subtasks: one of misogynistic tweet detection, the second of classification of the misogynistic tweets into the categories designed by (Anzovino et al., 2018) and into target types (against an individual or a group). The SemEval 2019 task (Basile et al., 2019), dealt with detecting hateful language related to women and immigrants in English and Spanish, and involved different classifications: hateful speech identification, specific or generic target, and aggressive/non- aggressive hateful text. The AMI-2020 shared task misogynistic and aggressive language identification and unbiased mysogyny identification.

Among other datasets, the misogyny dataset for Danish recently created by Zeinert et al. (2021), is particularly relevant to our own work in that the annotators are distinguished by characteristics including gender, age, ethnicity, and study/ occupation. The authors designed a taxonomy for labeling misogyny which uses four labels from Anzovino et al. (2018) plus two new labels: "neosexism" (discrimination denial) and "benevolent sexism" (attitudes toward gender in a positive way). Parikh et al. (2019) Created a multi-label sexism dataset that include 23 categories: Role stereotyping, Attribute stereotyping, Body shaming, Hyper-sexualization, Internalized sexism, Pay gap, Hostile work environment, Denial or trivialization of sexist misconduct, Threats, Rape, Sexual assault, Sexual harassment, Tone policing, Moral policing, Victim blaming, Slut shaming, Motherhood-related discrimination, Menstruation-related discrimination, Religion-based sexism, Physical violence, Mansplaining, Gaslighting and Other.

To the best of our knowledge there is only one dataset for misogynistic language in Arabic, the Let-Mi Levantine Arabic misogyny dataset created by Mulki and Ghanem (2021) and covering the Twitter accounts of seven female journalists who covered the protests in Lebanon. The data were annotated according to a scheme including the six sub-categories used in AMI-2018, and added a new category, "damning," used in Arab culture–asking God to hurt a woman. Unlike the dataset proposed in this paper, Let-Mi, like other datasets for studying misogyny/sexism, does not encode the effect of annotators characteristics on judgments, and the cases of disagreement were resolved

[4] In other recent literature (Parikh et al., 2021) misogyny are defined in more restricted interpretation as 'hate or entrenched prejudices against women' and the term sexism is used as a more general term that includes discrimination or judging a person (women in particular) based on gender.

with traditional aggregation methods. In this paper we argue that misogyny and sexism annotation heavily depends on subjective criteria that lead different people to label the same text in a different way based on their beliefs, and that such disagreements due to subjectivity should not be solved with aggregation procedures.

2.3. Inter-coder (dis)agreement in offensive language annotation

The work on creating datasets for hate speech, misogyny and sexism discussed in the previous sections focused on creating gold standard datasets either through manual reconciliation or through automatic aggregation (Ide and Pustejovsky, 2017; Paun et al., 2021). However, there is increasing awareness that this approach is only appropriate for the very special case of labelling tasks in which coders can be expected to agree on all or most cases (Uma et al., 2021; Basile, 2020). Intuitively, this assumption does not hold for subjective annotation tasks; this intuition is supported by the evidence on intercoder agreement on misogyny annotation. Anzovino et al. (2018) reported bias-adjusted kappa values of 0.4874 for misogyny identification and 0.3732 for misogyny categorization. In Evalita-2018 (Fersini et al., 2018a) the inter-annotator agreements for misogyny was 0.45 for English and 0.68 for Italian . In the SemEval-2019 shared task (Basile et al., 2019), the inter-rater agreement for aggressiveness in Spanish is 0.47. Zeinert et al. (2021), obtained 0.54 agreement between annotators and used traditional methods to mitigate biases in annotation such as several rounds of discussion and revision with annotators.

However, disagreement is not always low. In AMI Evalita-2018 (Fersini et al., 2018a) a high inter-rater annotator agreement was found for misogyny identification: 0.81 for English and 0.96 for Italian. In SemEval-2019 (Basile et al., 2019), the inter-rater agreement for English was 0.83 for hate and 0.73 for aggressiveness. Mulki and Ghanem (2021) obtained 5,529 tweets with Unanimous agreement, 1,021 tweets with mild agreement (two out of three) and 53 tweets with conflicts from three annotators; the inter-annotator agreement, computed using Krippendorff's alpha, was 82.9%.

These inconsistent results suggest that (i) the assumption of agreement among annotators does not clearly hold for misogyny and sexism, but (ii) the disagreements on misogyny and sexism judgments may be due to differences in subjective judgments, so that low agreement is found if the coders have different subjective biases, whereas high agreement can result if their biases match. (A similar hypothesis was made by Al Kuwatly et al. (2020), who argued that demographic features of a coder such as age, education and first language that can affect their individual judgments and may impact hateful language classifier models should be encoded.) This hypothesis led us to explore in this work the effect of subjectivity in misogyny and sexism annotation (for Arabic). In the next Sections (3.3 and 3.4) we report experiments showing that Arabic misogyny and sexism labelling is indeed a highly subjective task, where different annotators may have different interpretations based on their background beliefs–in particular, whether they are conservative or liberal.

2.4. Disagreement and gold standards

The view that the gold standard assumption is not appropriate for misogynous and sexist language, or indeed for other tasks depending on subjective judgments, is increasingly accepted for subjective tasks (Akhtar et al., 2019; Akhtar et al., 2020; Basile, 2020; Leonardelli et al., 2021). In fact, it is finding more and more adherents for other NLP tasks as well (Poesio and Artstein, 2005; Plank et al., 2014b; Aroyo and Welty, 2015; Peterson et al., 2019; Poesio et al., 2019; Uma et al., 2020; Fornaciari et al., 2021; Uma et al., 2021).

Aroyo and Welty (2015) argued that there are seven misconceptions concerning manual annotations that affect the collection and annotation of data: (i) there is a single correct data annotation, (ii) annotators should not disagree, (iii) annotators should be guided in detail to avoid disagreements, (iv) annotation is better done by experts, (v) it is better to have the evaluation by a single expert annotator, (vi) all samples are treated in the same way and (vii) annotations need no updating. More recent arguments for these positions can be found in (Basile, 2020; Basile et al., 2021).

In the case of abusive language, Akhtar et al. (2019) show how the coders characteristics can impact the judgments in the hate speech task, introducing an automated method for partitioning a set of coders into categories based into their backgrounds by measuring the polarization of all their judgments on each item. Klenner et al. (2020) show how 'harmonization sometimes harms' when producing a gold standard for a high subjective task (sentiment analysis).

As a result, a number of NLP authors including, e.g., (Poesio and Artstein, 2005; Poesio et al., 2013; Plank et al., 2014b; Aroyo and Welty, 2015; Poesio et al., 2019; Dumitrache et al., 2019; Pavlick and Kwiatkowski, 2019; Peterson et al., 2019; Basile, 2020; Uma et al., 2021) argued for datasets in which all annotations are preserved instead of creating a possibly artificial gold standard.

2.5. Learning from Annotator Distributions

The move towards datasets in which all judgments are preserved naturally leads to the question of how such datasets can be used to train NLP models. Akhtar et al. (2020) used the method from (Akhtar et al., 2019) to partition the coders in two groups based on their judgments to produce different group-based gold standard datasets; then each of these datasets was used to train a separate classifier to model the group perspective.

But in recent years there has been a lot of work on learning from annotator disagreement without resolving hard disagreement cases (Uma et al., 2021). Well-known proposals include (Sheng et al., 2008; Plank et

al., 2014a; Guan et al., 2018; Rodrigues and Pereira, 2018; Peterson et al., 2019; Uma et al., 2020; Fornaciari et al., 2021). As an alternative to the multiple-models approach proposed by Akhtar et al. (2020), we focus here on the **soft loss** approach proposed in (Peterson et al., 2019; Uma et al., 2020), which was found by Uma et al. (2021) to perform best with datasets for which a gold standard could not be defined.

In the soft-loss approach, a model is trained on the full **soft label**–a probability distribution derived from the crowd annotations–minimizing a loss function such as cross-entropy that does not require the target to be a one-hot label. Such models aim at mimicking crowd uncertainty, by capturing the full distribution over labels based on annotator confusion among the categories. Uma et al. (2020) tested the approach on a variety of datasets with different crowd characteristics, showing that the best method for generating the soft label (Softmax or standard normalization), depends on the number of annotators in the dataset and the annotation characteristics. Another novelty in (Peterson et al., 2019; Uma et al., 2020; Uma et al., 2021) was to evaluate the models not only by using hard metrics such as accuracy, but also soft metrics, such as cross entropy, capturing which model produce distribution most like the crowd's.

3. Data collection and annotation

The first contribution of this work is the creation of a dataset, the ArMIS corpus, to study subjectivity in misogyny and sexism data, and how such a dataset could be used in misogyny and sexism detection. In this Section we discuss the methods used to create the corpus and further annotation studies we carried out to investigate questions raised by the annotation.

3.1. Collection and preprocessing

In October 2020, we collected 2K Arabic misogyny and sexism tweets via the Twitter API, using a keywords list manually created specifically for this task, including specific slang words, phrases and hashtags, to get related tweets. We also used general terms e.g., 'women' and 'girls' to avoid the effect of biased keywords in downloading. Some examples of keywords we used are shown in Table 1. The Twitter API feature of 'extended' text was used to get the full length of the posted tweets.

The resulting collection contains tweets in a variety of Arabic dialects. We only removed duplicated tweets, non-Arabic text, advertisement, user mentions, retweets, and URLs. In Figure 1 we present a visualization of the most frequent words in ArMIS, and in Table 2 the top 5 most frequent words.

3.2. Annotation Scheme

3.2.1. Previous Annotation Schemes

As discussed in Section 2, a number of corpora annotated for misogyny and sexism exist. One option would have been the widely used annotation scheme developed for the SemEval-2019 task annotation scheme on detecting hate speech towards women and immigrants (Basile et al., 2019), in which the tweets are labelled as "hate" or "not hate". However, as already discussed in Section 2, labelling misogyny / sexism is not the same as labelling hate towards women, although such hate would be considered misogynistic. For example, in our data we find tweets such as Example 1, which does not express hate towards women (if at all, it expresses hate towards the man in question), but can nevertheless be considered misogynistic in the extended sense used by Fersini et al.

- Example 1: " البنت مفيش عتاب عليها عشان البنات ناقصات عقل ودين بس صاحبك ده عايز الحرق برجوله "

  "The girl is not blamed because girls are defect of mind and religion, but this guy feet should be burned"

Keyword	Keyword
ناقصات *deficient/imperfect*	متبرجات *dressed up*
نسويات *feminists*	فسويات *fart-eminist*
ملعونات *cursed*	ساقطات *bitches*
عاهرات *prostitutes*	فاسدات *corrupt*
فاجرات *whores*	متمردات *rebels*
صايعات *players*	فاسقات *sluts*
متحررات *liberated*	رخيصات *cheap*
مفسدات *spoilers*	سافرات *face revealing*
متسلطات *bossy*	سافلات *varmint*

Table 1: Some of the Arabic keywords used for data collection and their English translation

Figure 1: Visualization of most frequent Arabic words in ArMIS

Word	Count
النساء *women*	201
الله *God*	170
المرأة *woman*	160
ناقصات *deficient/imperfect*	132
عقل *mind*	132

Table 2: The top 5 most frequent terms in ArMIS

Since we are focusing on general toxic language towards women, using the labels ”misogyny” and ”not misogyny” seemed most appropriate, where ”misogyny” covers any type of misogynistic text including hate, aggressiveness and offensive language towards women (Poletto et al., 2021; Anzovino et al., 2018).

3.2.2. The ArMIS Annotation Scheme

Thus, for our annotation we used a slightly revised version of the scheme from AMI-2020 at Evalita (Fersini et al., 2020). This is a binary classification scheme with two labels: 1 for ”misogyny”, 0 for ”not misogyny”, as exemplified in Table 3. As discussed before, misogyny as defined by Fersini et al is more general than simply hate against women, also covering what in other schemes would be called sexist speech. The annotators were given the instructions provided in AMI-2018 and were asked to choose a label based on their perspective, assigning the ”misogyny” label if a tweet falls into one of the six categories used in AMI-2018 (Fersini et al., 2018a; Fersini et al., 2018b). Otherwise, the tweet should be classified as ”not misogyny”.

3.3. Annotation

The first version of the corpus was created to investigate two factors that would appear to affect disagreement in misogyny and sexism annotation: gender and religious (Islamic) beliefs.

We selected 964 tweets among those collected as discussed above, and had each tweet annotated by three annotators – 2 female and 1 male who self defined their degree of Islamic beliefs as Liberal, Moderate and Conservative.

Using Fleiss' Kappa (Fleiss, 1971) we obtained an overall Kappa of 0.525 for the agreement between the three annotators. We obtained Kappa=0.572 for the comparison between two females with different beliefs-one moderate, the other more liberal. The agreement between a moderate female and conservative male was Kappa=0.552. Finally, the result of reliability between a male and a female annotator where the former is conservative, and the latter is liberal, was Kappa=is 0.444. (See Table 4 for a summary). These results are in line with the low agreement found in some of the previous studies, and support the hypothesis that this low agreement may be affected by subjective factors.

In order to test the dependency between annotator characteristics and judgments, we ran a series of χ^2 tests. A χ^2 test comparing two annotators differing in both gender and beliefs–i.e., Liberal Female vs Conservative Male– (see Table 5) is highly significant (p=1.17E-07 p-value), but it is not clear which factor plays a role. A second test based on beliefs only–one Female Liberal annotator, one Moderate (Table 6) is also highly significant (p= 1.70E-10), suggesting a dependency on annotator beliefs.

Since we did not have coders with same beliefs but different gender, we ran a test considering the annotations of the two coders with the closest beliefs, the Moderate Female and the Conservative Male. The results (in Table 7) show that there is no significance in this case, with p=.2721. This would appear to suggest that gender is less predictive than beliefs.

3.4. Further annotation experiments

To verify the hypotheses of dependencies between judgments and coder characteristic suggested by the first annotation we carried out a larger scale study, in which a larger number of participants with different beliefs and gender were asked to annotate a subset of the data. 11 particularly controversial tweets from ArMIS were selected. These tweets were annotated by 32 subjects according to the ArMIS scheme. The subjects included an equal number from both genders (F=16, M=16), and were asked to characterise themselves as Liberal, Moderate or Conservative, resulting in beliefs distribution of Liberal: F=3, M=2, Moderate: F= 13, M= 11, and Conservative: F=0, M=3. The percentages of annotators according to gender and beliefs are shown in Figure 2.

We used χ^2 tests to compute the dependency between judgments and annotator characteristics. The results (Table 8) show that there is dependency between the judgments and participants' beliefs. However, the results in Table 9 and Table 10 show that whereas the males' beliefs has significantly impact on their judgments, that was not the case for females. To assess the impact of gender, we divided the participant who have the same beliefs (eg., Liberal or Moderate) into two groups, female and male and run the χ^2 test based on that. The results in Table 11 and Table 12 further confirm that gender does not have a significant effect on judgments.

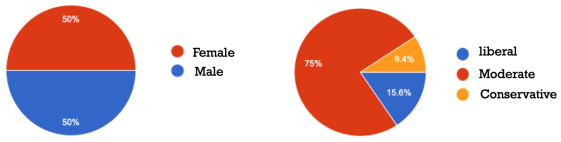

Figure 2: The distribution of 32 annotators based on gender and beliefs

3.5. The correlation between annotator characteristics and judgments

The results from our studies provide further evidence that annotating misogyny and sexism is a subjective task, where judgments can be influenced by the coder's beliefs. Although we do not expect that all annotators with the same characteristic will agree on all judgments, still the fact that the beliefs of coders have a high impact on their judgments cannot be ignored. A model of misogyny and sexism detection should learn all different perspectives, and should be able to produce labels that reflect different points of views instead of a single label in order that can be matched with a 'gold' label

Label	Instructions and examples
Misogyny	any text that expresses hating toward women in particular including discredit, sexual harassment, threats of violence, stereotype, objectification, derailing and dominance
	"ميهمناش لو برضه مشيتي على كواكب مجرة درب التبانه كده كوكب كوكب برضه ناقصات عقل و دين" *"We don't care even if you walk on the planets of the Milky Way, planet by planet, women are still defect of reason and religion"*
Not Misogyny	a text that does not express hating towards women in particular
	"زمن النساء شكرا محمد سلمان" *"Women time thank you Mohammed Salman"*

Table 3: Instructions for misogyny identification from (Fersini et al., 2018a; Fersini et al., 2018b) and examples from ArMIS

	Fleiss Kappa
Overall	0.525
$\text{MOD}_F - vs - LIB_F$	0.572
$\text{MOD}_F - vs - CON_M$	0.552
$\text{LIB}_F - vs - CON_M$	0.444

Table 4: Agreement between three annotators with different beliefs and gender.

	LIB_F	CON_M
Misogyny	311	424
Not misogyny	653	540
P-value:	**0**	

Table 5: χ^2 test between Liberal female and Conservative male annotators based on two factors gender and beliefs

	LIB_F	MOD_F
Misogyny	311	448
Not misogyny	653	516
P-value:	**0**	

Table 6: χ^2 test between two females annotators based on beliefs

	MOD_F	CON_M
Misogyny	448	424
Not misogyny	516	540
P-value:	.2721	

Table 7: χ^2 test between female and male annotators with the most close beliefs, based on gender

	LIB	MOD	CON
Misogyny	44	190	13
Not misogyny	11	74	20
P-value:	**.0001**		

Table 8: χ^2 test between 32 annotators based on beliefs

	LIB_M	MOD_M	CON_M
Misogyny	16	88	13
Not misogyny	6	33	20
P-value:	**.0013**		

Table 9: χ^2 test between 16 males based on beliefs

	LIB_F	MOD_F
Misogyny	28	102
Not misogyny	5	41
P-value:	.1111	

Table 10: χ^2 test between 16 females based on beliefs

	LIB_F	LIB_M
Misogyny	28	16
Not misogyny	5	6
P-value:	0.2709	

Table 11: χ^2 test between 5 annotators based on gender with same beliefs: Liberal

	MOD_F	MOD_M
Misogyny	102	88
Not misogyny	41	33
P-value:	0.801	

Table 12: χ^2 test between 24 annotators based on gender with same beliefs: Moderate

based on some form of aggregation. To achieve this, a misogyny and sexism dataset should include the views of coders with different relevant characteristics. The evidence presented in this paper suggests that religious beliefs is one such characteristic; we intend to explore more in subsequent work.

3.6. The ArMIS corpus

In summary, ArMIS contains 964 Arabic tweets annotated by three coders who self defined their degree of religious beliefs. 11 of these tweets have additional an-

notations from 32 annotators also who self defined their degree of religious beliefs (and gender). The dataset will be made publically available on github.

4. Learning misogyny and sexism detection models from disaggregated data

A second claim is that creating an artificial gold label is not necessary in order to train models for misogyny and sexism detection. In this section we discuss experiments in which we used ArMIS to show that it is possible to train misogyny and sexism detection models using data with disagreement and without the use of aggregated data. In these experiments we employed two different approaches to learning from disagreement: the soft loss training approach of (Peterson et al., 2019; Uma et al., 2020) and the 'multiple classifiers' approach of (Akhtar et al., 2019; Akhtar et al., 2020; Basile, 2020).

4.1. Using ArMIS for modelling

We split the 964 tweets into 674 for training, 145 for validation and 145 for testing. Given the nature of the task, we did not attempt to establish 'gold' labels; we did however use majority voting to produce a hard label for hard evaluation purposes and also to train the base model for the purpose of comparison.

4.2. Soft Loss Training

4.2.1. Methodology

To use the soft loss training approach proposed in (Peterson et al., 2019; Uma et al., 2020), we trained soft loss function using AraBERT models (Antoun et al., 2020) with soft labels as a target. This to minimize the loss between the Softmax probability distribution produced by a model and soft label distributions for ArMIS derived from the annotations as discussed next.

4.2.2. Generating the soft labels

We tested two ways to generate the soft labels for each item. The first approach involved using a standard normalization function (Peterson et al., 2019): for each training item, the probability of a label is the number of annotators who have chosen that label divided by total number of annotators. In the second approach, the soft labels were generated by using in addition a Softmax as proposed in (Uma et al., 2020).

The soft labels produced using Softmax or standard normalization were then used as targets for training using a soft loss function such as Cross Entropy which minimizes the loss between the probability distribution produced by the model and the soft label. The result in Table 13 suggest that standard normalization works best with ArMIS according to all metrics; however the difference is not significant.

4.2.3. Comparing soft loss functions

In addition to Cross Entropy we also tried other soft loss functions proposed in (Uma et al., 2021), including Mean-squared error MSE, Forward KL-divergence and Reverse KL-divergence. We used each of these functions to train the state-of-the-art AraBERT base model (Antoun et al., 2020) on soft labels. We used a maximum sequence length 128, learning rate 1e-5, batch size 8, and training for 10 epochs. All the reported results are based on an average of 10 runs for each model.

The results of this comparison are reported in Table 14. Using Cross Entropy as a loss function yields the best Accuracy and JSD (Table 14); however, training with MSE achieves the best results in terms of CE. None of the differences in Accuracy are significant, but the difference in CE results between the best soft loss function and the worst (Forward KL) are significant.

4.3. Hard training of separate classifiers

An alternative approach proposed by (Akhtar et al., 2019; Akhtar et al., 2020; Basile, 2020) is to train a separate classifier for each coder. We trained three AraBERT models, one for each coder, using Cross Entropy with one hot encoding. The accuracy with which these classifiers model the three annotators' perspectives is illustrated in Table 15. We can then use the outputs of these three classifiers to compute a hard label for each item using Majority vote, as well as a soft label using either standard normalization or Softmax as discussed above for soft loss training.

4.4. Majority vote hard training

Our baseline model was obtained using Majority vote to train a AraBERT classifier.

4.5. Hard and soft evaluation

Another issue to be addressed when using datasets providing multiple labels for each item is how to evaluate a model. It would make little sense to evaluate a misogyny and sexism detection model against a gold label except perhaps if the gold label is meant to capture the 'majority point of view'. For this reason, (Uma et al., 2021; Basile et al., 2021) argued that soft evaluation metrics are more appropriate at least for subjective tasks. In this work we used two soft evaluation metrics comparing the distance between probability distributions: Cross Entropy (CE), as used in (Peterson et al., 2019; Uma et al., 2020; Uma et al., 2021) and Jensen-Shannon Divergence (JSD), a symmetric version of Kullback-Leibler divergence (Uma et al., 2021). We also produced hard labels using majority voting for evaluation using Accuracy and F1, but only for comparison purposes.

4.6. Results

The results achieved by the three training approaches are reported in Table 16. All results are based on an average of ten runs for each model.

Training separate classifiers achieves better results in terms of the hard metrics–Accuracy and F1–as well as in terms of JSD. However, the best results in terms of Cross-Entropy are achieved using soft-loss training. These results confirm the findings e.g., of (Uma et al.,

2021) that whereas hard metrics tend to reward training with hard labels, the best results with soft metrics–which are arguably more appropriate for subjective tasks–are obtained with soft label training methods.

Model	ACC	F1	CE	JSD
$CE_{standard_norm}$	**77.79**	**77.38**	**0.586**	0.244
$CE_{Softmax}$	76.41	76.06	0.598	**0.194**

Table 13: CE comparison between soft labels generation using standard normalization and Softmax

Model	ACC	F1	CE	JSD
CE	**77.79**	**77.38**	0.586	**0.244**
MSE	76.83	76.09	**0.571***	0.261
Forward KL	76.41	75.80	0.942	0.251
Reverse KL	75.59	75.03	0.594	0.259

Table 14: Comparison between Soft loss functions using standard normalization soft labels as targets

Model	ACC	F1	CE	JSD
Liberal	**76.34**	**75.56**	0.886	**0.253**
Moderate	73.24	73.38	0.850	0.270
Conservative	73.59	73.53	**0.789**	0.269

Table 15: Separate annotators classifiers hard training on each annotator labels vs Majority vote

Model	ACC	F1	CE	JSD
CE soft loss	77.79	77.38	**0.586***	0.244
MV	76.89	76.42	0.906	0.245
Three classifiers	**78.00**	**77.67**	3.662	**0.205**

Table 16: Comparison between soft-loss training, hard training a single classifier using Majority vote, and training separate classifiers for each annotator

5. Related Work

The closest work to ours is the recent paper by Leonardelli et al. (2021), who created three datasets with balanced configuration with respect to domain topic, agreement level (A++,A+,A0) and label distribution. The authors used an ensemble of five different classifiers to annotate the data in order to select data to be annotated (each tweet annotated by 5 crowdworker) with the binary labels "offensive" or "not offensive". Each classifier was trained and evaluated on differently balanced sets. Their results show that including different levels of agreement in training and testing set impacts the classifier performance. When adding (A0) for training and test the model performance decreased, indicating that a task include ambiguity data is challenging to classify. Leonardelli et al also showed that high performance in the Offenseval shared task (Zampieri et al., 2020) is due to high agreement.

6. Conclusions and Future Work

In this paper we presented evidence in support of the claim that the observed disagreements in misogyny and sexism detection datasets are due to differences in coding characteristics. We argued that datasets for this task should include judgments from annotators reflecting these different characteristics, and introduced one such dataset, the freely available ArMIS corpus of Arabic misogyny and sexism. We also argued that the existence of such disagreements indicates the need for new approaches to training models for the task not based on the gold standard assumption, and of new soft metrics for evaluating such models. We provided proof-of-concept examples of how ArMIS and similar datasets can be used to train models for the task, and how such models can be evaluated. We are in the process of creating a larger version of the dataset and we are expanding the number of annotators with the use of a crowdsourced platform which involves introducing a more formal test to identify the degree of annotators religious beliefs, and select a balance number of annotators among different level of beliefs. this will allow us to clearly explore the effect of beliefs on misogyny and sexism annotation.

7. Acknowledgements

Dina Almanea is supported by the Saudi Arabian Cultural Bureau in the UK and the University of Jeddah in Saudi Arabia. Massimo Poesio was in part supported by the DALI project, ERC Advanced Grant 695662 to Massimo Poesio.

8. Bibliographical References

Akhtar, S., Basile, V., and Patti, V. (2019). A new measure of polarization in the annotation of hate speech. In *International Conference of the Italian Association for Artificial Intelligence*, pages 588–603. Springer.

Akhtar, S., Basile, V., and Patti, V. (2020). Modeling annotator perspective and polarized opinions to improve hate speech detection. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 151–154.

Al Kuwatly, H., Wich, M., and Groh, G. (2020). Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190.

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

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, May. European Language Resource Association.

Anzovino, M. E., Fersini, E., and Rosso, P. (2018). Automatic identification and classification of misogynistic language on twitter. In *NLDB*.

Aroyo, L. and Welty, C. (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36:15–24, 03.

Bartlett, J., Norrie, R., Patel, S., Rumpel, R., and Wibberley, S. (2014). Misogyny on twitter.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., Rosso, P., and Sanguinetti, M. (2019). SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Basile, V., Fell, M., Fornaciari, T., Hovy, D., Paun, S., Plank, B., Poesio, M., and Uma, A. (2021). We need to consider disagreement in evaluation. In *Proc. of the ACL-IJCNLP Workshop on Benchmarking: Past, Present and Future*.

Basile, V. (2020). It's the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks. In *Proc. of the AIXIA Workshop*. Universitá di Torino.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Dumitrache, A., Aroyo, L., and Welty, C. (2019). A crowdsourced frame disambiguation corpus with ambiguity. In *Proc. of NAACL*.

Fersini, E., Nozza, D., and Rosso, P. (2018a). Overview of the evalita 2018 task on automatic misogyny identification (ami). In *EVALITA@CLiC-it*.

Fersini, E., Rosso, P., and Anzovino, M. E. (2018b). Overview of the task on automatic misogyny identification at ibereval 2018. In *IberEval@SEPLN*.

Fersini, E., Nozza, D., and Rosso, P. (2020). Ami@ evalita2020: Automatic misogyny identification. In *EVALITA*.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Fornaciari, T., Uma, A., Paun, S., Plank, B., Hovy, D., and Poesio, M. (2021). Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proc. of NAACL*. Association for Computational Linguistics.

Frenda, S., Ghanem, B., y Gómez, M. M., and Rosso, P. (2019). Online hate speech against women: Automatic identification of misogyny and sexism on twitter. *J. Intell. Fuzzy Syst.*, 36:4743–4752.

Guan, M., Gulshan, V., Dai, A., and Hinton, G. (2018). Who said what: Modeling individual labelers improves classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Hewitt, S., Tiropanis, T., and Bokhove, C. (2016). The problem of identifying misogynist language on twitter (and other online social spaces). In *Proceedings of the 8th ACM Conference on Web Science*, pages 333–335.

Nancy Ide et al., editors. (2017). *The Handbook of Linguistic Annotation*. Springer.

Klenner, M., Göhring, A., Amsler, M., Ebling, S., Tuggener, D., Hürlimann, M., and Volk, M. (2020). Harmonization sometimes harms.

Leonardelli, E., Menini, S., Aprosio, A. P., Guerini, M., and Tonelli, S. (2021). Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. *arXiv preprint arXiv:2109.13563*.

Mubarak, H., Darwish, K., and Magdy, W. (2017). Abusive language detection on arabic social media. In *Proceedings of the first workshop on abusive language online*, pages 52–56.

Mubarak, H., Rashed, A., Darwish, K., Samih, Y., and Abdelali, A. (2020). Arabic offensive language on twitter: Analysis and experiments. *arXiv preprint arXiv:2004.02192*.

Mulki, H. and Ghanem, B. (2021). Let-mi: An arabic levantine twitter dataset for misogynistic language. *arXiv preprint arXiv:2103.10195*.

Mulki, H., Haddad, H., Ali, C. B., and Alshabani, H. (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.

Nockleby, J. T. (2000). Hate speech. *Encyclopedia of the American constitution*, 3(2):1277–1279.

Parikh, P., Abburi, H., Badjatiya, P., Krishnan, R., Chhaya, N., Gupta, M., and Varma, V. (2019). Multi-label categorization of accounts of sexism using a neural framework. *arXiv preprint arXiv:1910.04602*.

Parikh, P., Abburi, H., Chhaya, N., Gupta, M., and Varma, V. (2021). Categorizing sexism and misogyny through neural approaches. *ACM Transactions on the Web (TWEB)*, 15(4):1–31.

Paun, S., Artstein, R., and Poesio, M. (2021). *Statistical Methods for Annotation Analysis*. Morgan Claypool.

Pavlick, E. and Kwiatkowski, T. (2019). Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Peterson, J. C., Battleday, R. M., Griffiths, T. L., and Russakovsky, O. (2019). Human uncertainty makes

classification more robust. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9616–9625.

Plank, B., Hovy, D., and Sogaard, A. (2014a). Learning part-of-speech taggers with inter-annotator agreement loss. In *Proc. of EACL*.

Plank, B., Hovy, D., and Søgaard, A. (2014b). Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511.

Poesio, M. and Artstein, R. (2005). The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In A. Meyers, editor, *Proc. of ACL Workshop on Frontiers in Corpus Annotation*, pages 76–83, June.

Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L., and Ducceschi, L. (2013). Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Transactions on Intelligent Interactive Systems*, 3(1).

Poesio, M., Chamberlain, J., Kruschwitz, U., Paun, S., Uma, A., and Yu, J. (2019). A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. In *Proc. of NAACL*, page 1778–1789, Minneapolis. Association for Computational Linguistics (ACL).

Poland, B. (2016). *Haters: Harassment, abuse, and violence online*. U of Nebraska Press.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., and Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. *Lang. Resour. Evaluation*, 55:477–523.

Rodrigues, F. and Pereira, F. (2018). Deep learning from crowds. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Sheng, V. S., Provost, F., and Ipeirotis, P. G. (2008). Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 614–622.

Uma, A., Fornaciari, T., Hovy, D., Paun, S., Plank, B., and Poesio, M. (2020). A case for soft-loss functions. In *Proc. of HCOMP*.

Uma, A., Fornaciari, T., Hovy, D., Paun, S., Plank, B., and Poesio, M. (2021). Learning from disagreements. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Waseem, Z., Davidson, T., Warmsley, D., and Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*.

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., and Çöltekin, Ç. (2020). Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). *arXiv preprint arXiv:2006.07235*.

Zeinert, P., Inie, N., and Derczynski, L. (2021). Annotating online misogyny. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3181–3197.