Dina Almanea [1] Massimo Poesio[1]

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

[1]Queen Mary University of London, School of Electronic Engineering and Computer Science

Introduction

- 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.
- Few benchmarks for Arabic misogyny and sexism detection exist, and annotations are in aggregated form.
- However, misogyny and sexist judgments appears to depend on certain characteristics of annotators.
- We investigated 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 (ArMIS) characterized by providing annotations from annotators with different gender and different degree of Islamic 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 religious beliefs on misogyny and sexism annotation.
- 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 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.

Misogyny and Sexism

- 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).
- Misogyny is defined in (Parikh et al., 2021) in a more restricted way 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.
- Unlike the ArMIS dataset proposed in this paper, other shared tasks and datasets for studying misogyny/sexism do not encode the effect of annotators characteristics on judgments, and the cases of disagreement were resolved with traditional aggregation methods.
- In this paper we show evidence 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 religious beliefs, and argue that such disagreements due to subjectivity should not be solved with aggregation procedures.

Inter-coder (Dis)agreement in Misogynistic Language Annotation

- Gold standards are generally not appropriate for subjective annotation tasks.
- The inconsistent results on inter-coder agreement on misogyny annotation suggest that
  - The assumption of agreement among annotators does not clearly hold for misogyny and sexism, but
  - Low agreement might be found if the coders have different subjective biases, whereas high agreement can result if their biases match.
- Similar hypothesis in (Al Kuwatly et al., 2020), that demographic features of a coder can affect their individual judgments and may impact hateful language classifier models should be encoded.

ArMIS: Data Collection and Preprocessing

- 2K Arabic misogyny and sexism tweets were collected in October 2020 via the Twitter API.
- The resulting collection contains tweets in a variety of Arabic dialects.
- Duplicated tweets, non-Arabic text, advertisement, user mentions, retweets, and URLs were removed.

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*

- The top 5 most frequent words.

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

The ArMIS Annotation Scheme

- The binary classification scheme that was used is a slightly revised version of the scheme from AMI-2020 at Evalita (Fersini et al., 2020).
- Misogyny as defined by Fersini et al is more general than simply hate against women, and also covers what in other schemes would be called sexist speech.
- The annotators were asked **to choose a label based on their perspective**.

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"*

Annotation and the correlation between annotator characteristics and judgments

- First experiment: 964 tweets were selected, and annotated by three annotators
  - Liberal female, Moderate female and Conservative male.
- Second experiment: 11 controversial tweets from ArMIS were selected and annotated by 32 annotators.
  - gender: (Female=16, Male=16)
  - religious beliefs:(Liberal: F=3, M=2), (Moderate: F= 13, M= 11), and (Conservative: F=0, M=3).

Three annotators (964 tweets)

	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

32 annotators (11 controversial tweets)

	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 belief

	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

Using ArMIS for Modelling

- 964 tweets divided into 674 for training, 145 for validation and 145 for testing.
- Majority voting was used only to produce a hard label for hard evaluation purposes and also to train the base model for the purpose of comparison.

Learning from Disagreement 1: Soft Loss Training (Peterson et al., 2019; Uma et al., 2020)

- Soft loss function was trained using AraBERT models (Antoun et al., 2020) with soft labels as a target.
- Soft labels generation 1 : standard normalization function (Peterson et al., 2019).
- Soft labels generation 2 : Softmax as proposed in (Uma et al., 2020).
- The soft labels produced were then used as targets for training using separate soft loss function such as Cross Entropy.

Learning from Disagreement 2: Hard Training of Separate Classifiers (Akhtar et al., 2019, 2020; Basile, 2020)

- Training three separate classifier for each coder using AraBERT models on Cross Entropy with one hot encoding.
- The outputs of the three classifiers were used to compute a hard label for each item using Majority vote, and soft label using either standard normalization or Softmax.

Hard and soft evaluation

- It would make little sense to evaluate a misogyny and sexism detection model against a gold label.
- Two soft evaluation metrics were used for comparing the distance between probability distributions:
  - Cross Entropy (CE) (Peterson et al., 2019; Uma et al., 2020, 2021)
  - Jensen-Shannon Divergence (JSD), a symmetric version of Kullback-Leibler divergence (Uma et al., 2021).
- majority voting labels were produced for hard evaluation Accuracy and F1, but only for comparison purposes.

Results

- The results suggest that standard normalization works best with ArMIS according to all metrics.

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

- Hard metrics tend to reward training with hard labels.
- The best results achieved with soft metrics 'Cross-Entropy' using soft-loss training, which are more appropriate for subjective tasks.

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**

Acknowledgments

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.

References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 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.

Sohail Akhtar, Valerio Basile, and Viviana Patti. 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.

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 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.

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, Marseille, France. European Language Resource Association.

Mary E. Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *NLDB*.

Valerio Basile. 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.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2020. Ami@ evalita2020: Automatic misogyny identification. In *EVALITA*.

Pulkit Parikh, Harika Abburi, Niyati Chhaya, Manish Gupta, and Vasudeva Varma. 2021. Categorizing sexism and misogyny through neural approaches. *ACM Transactions on the Web (TWEB)*, 15(4):1–31.

Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. 2019. Human uncertainty makes classification more robust. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9616–9625.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2020. A case for soft-loss functions. In *Proc. of HCOMP*.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreements. *Journal of Artificial Intelligence Research*, 72:1385–1470.