Introduction

- The use of misogynistic and sexist language has increased in recent years in social media, and is increasing in the Arab world in reaction to reforms attempting to impose restrictions on women’s lives.
- Few benchmarks for Arabic misogyny and sexism detection exist, and annotations are in aggregated form.
- However, misogyny and sexist judgments appear 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).

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 behaviors like social exclusion and discrimination to more severe expressions related to intimacy, violence and sexual objectification” (Anzovino et al., 2018).
- Misogyny is defined in (Parkhi et al., 2021) in a more restricted way: “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 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.

The top 5 most frequent words.

<table>
<thead>
<tr>
<th>Word</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>women</td>
<td>1076</td>
</tr>
<tr>
<td>men</td>
<td>933</td>
</tr>
<tr>
<td>islam</td>
<td>743</td>
</tr>
<tr>
<td>say</td>
<td>337</td>
</tr>
<tr>
<td>time</td>
<td>259</td>
</tr>
</tbody>
</table>

The ArMIS Annotation Scheme

- The binary classification scheme that was used in a slightly revised version of the scheme from AMI-2020 at Eskita (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.

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

- Soft loss function was trained using AMI@MIL 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 (Umra et al., 2020).
- The soft labels produced were then used as targets for training using separate soft loss function such as Cross-Entropy.

Results

- The results suggest that standard normalization works best with ArMIS models on Cross-Entropy with one hot encoding.
- The outputs of the three classifiers were used to compute a hard label for each tweet using Majority vote, and soft label using either standard normalization or Softmax.

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

- Training three separate classifiers for each coder using stratified models on Cross-Entropy with one hot encoding.
- The results achieved with soft metrics ‘Cross-Entropy’ using soft-loss training, which are more appropriate for subjective tasks.

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; Umra et al., 2020).
  - Jensen-Shannon Divergence (JSD) (2005) of a symmetric version of Kullback-Leibler divergence (Umra et al., 2020).
- majority voting labels were produced for hard evaluation Accuracy and F1, but only for comparison purposes.

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 DALLI project, ERC Advanced Grant 695662 to Massimo Poesio.

References

- Massimo Poesio was in part supported by the DALI project, ERC Advanced Grant 695662 to Massimo Poesio.

\[ CE = - \sum_{i,j} p_{ij} \log q_{ij} \]