ASA CONFERENCE 2022
Data-Driven Decision Making
Genoa, 12-14 September 2022

Associazione per la Statistica Applicata
Applied Statistics Association

An experimental annotation task to investigate annotators' subjectivity in a misogyny dataset

Alice Tontodimamma, Stefano Anzani, Elisa Ignazzi and Lara Fontanella:

«Gabriele d'Annunzio» University, Chieti-Pescara

Marco Antonio Stranisci and Valerio Basile:

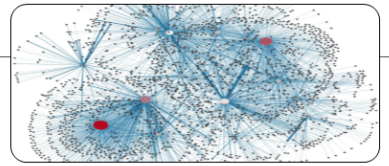
University of Turin

Outline

- Icomic Project
- Definition of Misogyny
- Misogyny Annotation Task
- Subjective Annotation tasks
- Research Aim
- Textual Corpora
- Annotation Task
- Inter Annotation Agreement
- Quantitative-qualitative analysis of disagreement
- Conclusion and Future Work

ICOMIC: Identifying and Countering Online Misogyny

Project funded by EU Next Generation, MUR-Fondo Promozione e Sviluppo-DM 737/2021



Misogynistic speech detection

O1.WP1

- Misogynistic Lexicon

O1.WP2

- **Annotated corpus for misogyny detection**

O1.WP3

- Machine learning models, exploiting lexical, lexicon-based, and sentiment analysis features, with XAI approaches

Identification of producers of misogynistic content

O2.WP1

- Twitter corpus (politically-active women, feminists, entrepreneurs, journalists, influencers, and celebrities)

O2.WP2

- Classification of misogynistic tweets and users

O2.WP3

- Ego-networks identification e representation

O2.WP4

- Nodes classification

O2.WP5

- Community detection and interpretation

Countering misogynistic speech

O3.WP1

- Literature systematic review

O3.WP2

- Expert consultation: Delphi methods

O3.WP3

- Expert consultation: Future scenarios

O3.WP4

- Expert consultation: Cross-impact

Definition of misogyny

- Merriam-Webster's online dictionary: Hatred of, aversion to, or prejudice against women. Misogyny refers specifically to a hatred of women.
- Oxford English Dictionary : A feeling of hate or dislike towards women, or a feeling that women are not as good as men.
- Australia's Macquarie Dictionary: 1. hatred of women. 2. entrenched prejudice against women.
- American Heritage Dictionary: Hatred or mistrust of women.
- Treccani online Italian dictionary: Atteggiamento di avversione generica per le donne.

Misogyny definition for the annotation task: *atteggiamento di avversione, mista a disprezzo e ostilità, verso le donne in generale*.

Misogyny annotation task

- The foundation of supervised machine learning is the development of excellent gold standard datasets for training and benchmarking.
- In Natural Language Processing supervised models typically start with corpus annotated by humans.
- Obtaining multiple annotator judgements on the same data instances is a common practice in order to improve the quality of final labels, obtained through an aggregation procedure.
- However, annotation for the study of highly subjective phenomena, like abusive and insulting language, is posing serious problems due to the subjectivity nature of the task (Basile, 2020).
- The notion of a '**single correct answer**' might fail to take into account the ***subjectivity*** and ***complexity*** of misogyny annotation task.

Subjective annotation tasks

- Human judgment applied to an "**objective**" task is wholly dependent on the item being judged. Ideally, different assessments should match and small differences in the annotations can be explained as measurement noise.
- Human judgment applied to a "**subjective**" task is **intrinsically influenced by factors pertaining to the judges themselves**.
  - Different people, while annotating a highly subjective task such as abusive language, can differ greatly in how offensive they find various expressions to be.

❖ ***In the subjective task scenario, the one-truth assumption is no longer valid*** (Basile, 2020).

In such cases, the opinions of ***all the annotators could be seen as valid****.*

❖ Proposals have been made to consider disagreement as an information content that can be exploited to improve the supervised classification performance (Basile et al., 2021).

Research aim

- **Assessment of misogyny annotation task subjectivity**
  - Exploration of annotators' agreement
  - Evaluation of the complexity of misogyny annotation task

The annotation process was carried by 12 trainees (2 males, 10 females, students enrolled on the Sociology degree course) who were engaged in an internship program in the Computational Social Research Lab.

Textual corpora

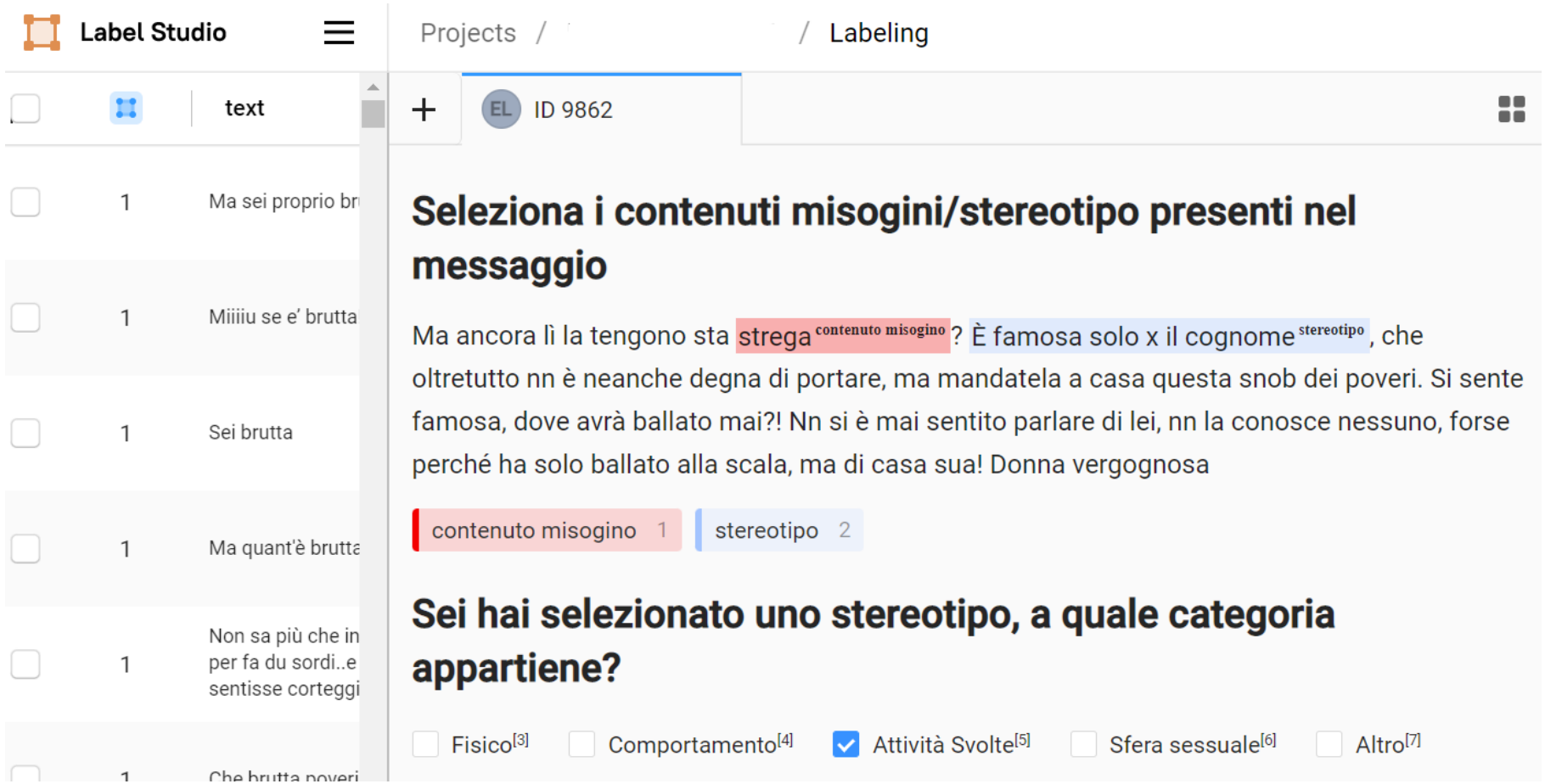
Twitter's Corpus:

- 760 messages posted on Twitter after the liberation of Silvia Romano on the 9th of May, 2020.
- Tweets were obtained through the official Twitter API and filtered by keywords: only messages published from the 9th to the 16th of May and containing the mention of Silvia Romano were collected and sampled.

Facebook's Corpus:

- 784 comments constructed starting from a total of 57826 Facebook comments to post directed to women and selected by the trainees themselves.
- These comments were scraped using **exportcomments.com**
- For the annotation task, we extracted a sample from this corpus using the revised HurtLex dictionary (Tontodimamma et al., 2022). We retained only comments containing words that belong to the revised HurtLex dictionary.

Annotation task



- Each comment is annotated by 4 annotators.

Inter Annotation Agreement

- Binary classification task *classes*$= \{M, \bar{M}\}$: **Confusion matrix**

		Annotator B	
		M	$\bar{M}$
Annotator A	M	TP	FN
	$\bar{M}$	FP	TN

- The computation of the Inter Annotation Agreement (IAA) relied on
  - **Cohen's Kappa** (Fleiss, 1969) for **labels**
  - **F1-measure** (Lehnert, 1992) for **spans**

Cohen's Kappa for a binary annotation

$$\kappa = \frac{p_0 - p_e}{1 - p_e}$$

- p_0 : proportion of observed agreement in the **labels** between two annotators $\sum_{i=0}^{1} p_{ii}$
- p_e: proportion of chance agreement $\sum_{i=0}^{1} p_{i.}\, p_{.i}$

Cohen's kappa is designed for measuring the agreement between two raters.

For multiple raters: average of the kappa statistics computed from each possible pair of raters.

$$-1 < k < 1$$

< 0.00 indicate poor agreement, 0.00–0.20 indicate slight agreement, 0.21–0.40 indicate fair agreement, 0.41–0.60 indicate moderate agreement, 0.61–0.80 indicate substantial agreement, and 0.81–1.00 indicate almost perfect agreement.

F1-measure

- The annotations of one annotator are used as the reference against which the annotations of the other annotator are compared.
- For each pair of annotators A and B we have 2 F1-measure, one relative to A on B and the other B on A.
- If A is the baseline,
- 1. **select all comments annotated with the misogyny label by A**
- 2. compute the confusion matrix as
- 3. compute
  - *B precision* = $\frac{TP}{TP+FP}$
  - *B recall* = $\frac{TP}{TP+FN}$

		Annotator B	
		M	$\bar{M}$
Annotator A	M	TP:# tokens selected by A & B	FN: # tokens selected by A but not B
	$\bar{M}$	FP: # tokens selected by B but not A	TN: # tokens not selected by A & B

Quantitative analysis of disagreement

- **F1-Score** is the harmonic mean of recall and precision:

$$F1 = \frac{2 \times recall \times precision}{precision + recall}$$

- The average F1-measure among all pairs of raters can be used to quantify the agreement among the raters.
- The higher the average F1-measure, the more the raters agree in the span selection.

Quantitative analysis of disagreement

The mean of Cohen's Kappa coefficients scored by each annotator and F1-measure.

		Twitter's Corpus	Facebook's Corpus
labels Cohen's Kappa	mean	0.228	0.210
	std	*0.120*	*0.090*
spans F1-measure	mean	0.232	0.299
	std	*0.070*	*0.190*

Qualitative analysis of agreement

Example of comments with more and less agreement for Twitter dataset

Source	Text	Chunk 1	Chunk 2
Silvia Romano	Cmq se sapevo che bastava andare in Somalia a trombarmi uno dal lavoro un po 'fumoso' x poi sposarlo ed essere 'trattata bene' x far sganciare a mio marito 4 mioni dallo stato italiano, altro che vacanze in Kenya... Anche io in Somalia andavo. Poi si divide eh? #SilviaRomanoAisha	*trombarmi uno dal lavoro un po 'fumoso' x poi sposarlo ed essere 'trattata bene'*	*trombarmi uno dal lavoro un po 'fumoso' x poi sposarlo ed essere 'trattata bene'*
Silvia Romano	Ha chiesto il corano. Si è convertita all'Islam. Torna in Italia con gli stessi abiti che indossano le donne islamiche. Abbiamo regalato milioni di euro a terroristi. E Conte e Di Maio l'hanno pure accolta a braccia aperte. Schifo. #SilviaRomano #LiveNoneLadUrso	*Conte e Di Maio l'hanno pure accolta a braccia aperte*	*Schifo.*

Qualitative analysis of disagreement

Example of comments with more and less agreement for Facebook dataset.

Source	Text	Chunk 1	Chunk 2
Facebook	Ma il tuo gommista ti ha gonfiato anche il lato B?Oltre le tette...chiedo per un'amica un po' sgonfia...	*Ma il tuo gommista ti ha gonfiato anche il lato B?Oltre le tette*	*Ma il tuo gommista ti ha gonfiato anche il lato B?Oltre le tette*
Facebook	Capra,capra,capra!!! NN TOCCARE LA SICILIA!!! Soprattutto noi siciliani!!! Cn moltissimi valori!!!Quelli che nn tieni tu'!!! GALLINA SPENNATA!!	*GALLINA SPENNATA*	*Capra,capra,capra!!*

Conclusion and Future Work

Conclusion

- In this work we **explored agreement among annotators** considering two corpora developed through an experimental annotation task .
- The analysis of annotations showed disagreement in both tasks.
- **Understanding the nature and sources** of the disagreements found in a dataset would thus appear to be an **essential prerequisite** if we are to properly harness disagreement in building machine learning models (Uma et al., 2021).

Conclusion

Why?

- Disagreement can be a result of subjectivity and complexity of task.
  - It could improve performance of tasks.
- Disagreement can be a result of annotators errors or problems with annotation scheme.
  - It introduces noise to the data.

- In our case-study disagreement depends both from annotators errors and subjectivity of annotation task. The extent of this disagreement varies depending on the complexity and genre of the task and dataset.

Future Work

- This result can be considered to create **guidelines** for the next annotation tasks (the annotator must be an expert, the annotator must be trained, the concept of misogyny must be well defined).
- Future work will focus on analyse how disagreement impacts on computational resources and try to integrate disagreement into modelling and evaluation.

ASA CONFERENCE 2022
Data-Driven Decision Making
Genoa, 12-14 September 2022

THANK YOU FOR YOUR ATTENTION

alice.tontodimamma@unich.it
s.anzani92@gmail.com
marcoantonio.stranisci@unito.it
valerio.basile@unito.it
elisa.ignazzi@studenti.unich.it
lara.fontanella@unich.it

References

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 2020 AIxIA Discussion Papers Workshop, AIxIA 2020 DP (Vol. 2776, pp. 31-40). CEUR-WS.

Basile, V., Fell, M., Fornaciari, T., Hovy, D., Paun, S., Plank, B., ... & Uma, A. (2021). We Need to consider disagreement in evaluation. In 1st Workshop on Benchmarking: Past, Present and Future (pp. 15-21). Association for Computational Linguistics.

Nozza, D., Volpetti, C., & Fersini, E. (2019). Unintended bias in misogyny detection. In Ieee/wic/acm international conference on web intelligence (pp. 149-155).

Davani, A. M., Díaz, M., & Prabhakaran, V. (2022). Dealing with disagreements: Looking beyond the majority vote in subjective annotations. Transactions of the Association for Computational Linguistics, 10, 92-110.

Fleiss, J. L., Cohen, J., & Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. Psychological bulletin, 72(5), 323.

Lehnert, W., Cardie, C., Fisher, D., McCarthy, J., Riloff, E., & Soderland, S. (1992). University of Massachusetts: MUC-4 test results and analysis. In Fourth Message Uunderstanding Conference (MUC-4): Proceedings of a Conference Held in McLean, Virginia, June 16-18, 1992.

Tontodimamma, A., Fontanella ,L., Anzani, S., Basile V. An Italian lexical resource for incivility detection in online discourses. Quality & Quantity. https://doi.org/10.1007/s11135-022-01494-7.

Uma, A., Fornaciari, T., Dumitrache, A., Miller, T., Chamberlain, J., Plank, B., ... & Poesio, M. (2021). Semeval-2021 task 12: Learning with disagreements. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021) (pp. 338-347). Association for Computational Linguistics.