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

Alice Tontodimamma, Stefano Anzani, Marco Antonio Stranisci, Valerio Basile, Elisa Ignazzi, Lara Fontanella

1. Introduction

In recent years, hatred directed against women has spread exponentially, especially in online social media, where the detachment resulting from being enabled to write without any obligation to reveal oneself directly allows people to feel greater freedom in the way they express themselves, and even to attack a chosen target without risk of being recognised or traced. Although this alarming phenomenon has given rise to many studies both from the viewpoint of computational linguistics and from that of machine learning, less effort has been devoted to analysing whether models for the detection of misogyny are affected by bias (Nozza et al., 2019).

During the last years, the problem of social bias in the field of Natural Language Processing (NLP) has been increasingly considered. Obtaining multiple annotator judgements on the same data instances is a common practice in NLP in order to improve the quality of final labels.

However, the fact that annotators are individuals obviously means that they have their own biases and values, and therefore are often likely to disagree with each other, especially when they are working on subjective tasks which involve detecting offensive language, misogynistic language, and hate speech. These disagreements can have a positive value, since they isolate subtleties in tasks of this kind that are obscured when annotations are combined to create a single ground truth (Davani et al., 2022).

In this work, we present two corpora: a corpus of messages posted on Twitter after the liberation of Silvia Romano on the 9th of May, 2020 and corpus of comments constructed starting from posts on Facebook that contained misogyny, developed through an experimental annotation task, to explore annotators' disagreement. In particular, we propose a qualitative-quantitative analysis of the resulting corpora.

2. Related work

The notion of a 'single correct answer' fails to take into account the subjectivity and complexity of many tasks. A task can be defined as 'subjective' when the human judgement is inherently influenced by factors pertaining to the judges themselves, rather than by the linguistic phenomenon, whereas human judgement applied to an 'objective' task depends solely on the object that is being judged. Different people, while annotating a highly subjective task such as offensive language, can differ greatly in how offensive they find various expressions to be: in such cases, the opinions of all the annotators could be seen as valid. In the subjective task scenario, the one-truth assumption is no longer valid (Basile, 2020).

In recent years, proposals have been made to consider disagreement as an information content that can be exploited to improve the performance of tasks (Basile et al., 2021). Uma et al (2020)

Alice Tontodimamma, University of Chieti-Pescara G. D'Annunzio, Italy, alice.tontodimamma@unich.it
Stefano Anzani, University of Chieti-Pescara G. D'Annunzio, Italy, s.anzani92@gmail.com, 0009-0000-5408-0104
Marco Antonio Stranisci, University of Turin, Italy, marcoantonio.stranisci@unito.it, 0000-0001-9337-7250
Valerio Basile, University of Turin, Italy, valerio.basile@unito.it, 0000-0001-8110-6832
Elisa Ignazzi, University of Chieti-Pescara G. D'Annunzio, Italy, elisa.ignazzi@studenti.unich.it
Lara Fontanella, University of Chieti-Pescara G. D'Annunzio, Italy, lara.fontanella@unich.it, 0000-0002-5441-0035

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Alice Tontodimamma, Stefano Anzani, Marco Antonio Stranisci, Valerio Basile, Elisa Ignazzi, Lara Fontanella, *An experimental annotation task to investigate annotators' subjectivity in a misogyny dataset*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-0106-3.49, in Enrico di Bella, Luigi Fabbris, Corrado Lagazio (edited by), *ASA 2022 Data-Driven Decision Making. Book of short papers*, pp. 281-286, 2023, published by Firenze University Press and Genova University Press, ISBN 979-12-215-0106-3, DOI 10.36253/979-12-215-0106-3

and Basile (2020) studied the impact of disagreement-informed data on the quality of NLP evaluation, and found it to be beneficial and providing complementary information for the quality of classification tasks. There are also authors in contrast with this approach: Bowman and Dahl (2021) recently proposed to study biases and artifacts in data to eliminate them; Beigman Klebanov et al. (2009) adopted a slightly softer stance, proposing to only evaluating on 'easy' instances. Basile et al. (2021) argue against this approach, based on the evidence about the prevalence of disagreement in NLP judgments. Removing the disagreement could lead to better evaluation scores, but fundamentally it hides the true nature of tasks. Furthermore, the reduction of noise in the data leads to a loss of information.

Our work contributes to the topic of investigating the impact of disagreement on computational resources by presenting an experimental annotation pipeline aimed at enhancing the subjectivity of annotators. Rather than being bound to a rigid set of labels, annotators were asked to label texts with an open-ended annotation, highlighting the portion of text that they considered to be misogynistic. This type of task had already been proposed, for example in Toxic Spans Detection, which is a task at SemEval 2021(Pavlopoulos et al., 2021). In fact, in Toxic Span Detection participants were asked to identify toxic spans, i.e., proportion of text that were responsible for the toxicity of the posts, when identifying such spans was possible.

3. Dataset creation and description

The dataset creation process involved trainees engaged in an internship program, who participated in two annotation tasks. They first annotated a corpus of 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.

For the second task, trainees labelled 784 Facebook comments. We started 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), an Italian lexicon of offensive, aggressive, and hateful words divided into 21 categories. Specifically, we used three categories: derogatory words, words related to prostitution, and words used to offend, insult, or denigrate women, which we consider could be used to create a subset. Using this filter, we retained only comments containing words that belong to these three categories and that occur at least 8 times. The final dataset for the annotation task comprises 784 comments.

4. Annotation task

For a given comment, the annotation procedure consists in selecting one or more chunk from each text that is regarded as misogynistic and establishing whether a gender stereotype is present. Each comment is annotated by at least three annotators in order to better analyse their subjectivity. The annotation process was carried by 13 trainees (2 males, 11 females, students on the Sociology degree course) who were engaged in an internship program in the Computational Social Research Lab[1].

5. Quantitative-qualitative analysis of disagreement

As a result of the annotation task, 2,207 annotations of tweets about Silvia Romano and 4,942 annotations of Facebook posts were collected. Each Facebook message obtained 3 annotations, while 4 annotations were provided for each Tweet.

[1] http://csrlab.unich.it/.

Since annotation tasks about abusive language are highly prone to subjectivity (Basile *et al*, 2021) and chunk selection tasks often result in significant disagreement, in this section a quantitative and qualitative analysis of disagreement is provided. The computation of the Inter Annotation Agreement (IAA) relied on Cohen's Kappa (Fleiss, 1969) for labels, and F1-measure (Lehnert, 1992) for spans.

Specifically, Cohen's kappa is designed for measuring the agreement between two raters and it is defined in the following way:

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

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

When multiple raters are considered, the kappa statistics computed from each possible pair of raters are averaged. Kappa has value 1 if there is perfect agreement between the raters, and value 0 if the observed agreement is equal to agreement expected by chance. Several authors have suggested interpretation or benchmark guidelines for values between 0 and 1. Landis and Koch (1977) proposed the following guidelines: 0.00 - 0.20 indicates slight agreement, 0.21-0.40 fair agreements, 0.41-0.60 moderate agreement, 0.61-0.80 substantial agreement, and 0.81-1.00 indicates almost perfect agreement.

The IAA on chunk selection was computed only on messages annotated with the same label and was computed through averaged pairwise F1-measure, which is the harmonic of precision and recall. In this setting, the annotations of one annotator are used as the reference against which the annotations of the other annotator are compared. 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.

Table 1 shows the IAA agreement for both labelling and span detection activities. Values are the average of Cohen's Kappa scores and F1-measures obtained by each annotator against the others who annotated the same part of the dataset. In order to account the differences between single annotators we also computed the standard deviation for all tasks and activities.

		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*

Table 1: Mean and standard deviation of Cohen's Kappa coefficients scored by each annotator and F1-measure.

From a general overview of Cohen's Kappa scores first emerges a low agreement in both tasks. Annotators averaged an agreement of 0.228 on the Silvia Romano's task, and of 0.210 on the Facebook posts task. It is worth mentioning the high standard deviation between annotators, which is 0.12 for the former task and 0.09 for the latter.

For the F1-measures results show that annotators obtained a higher agreement selecting span from Facebook posts than from tweets about Silvia Romano. However, the standard deviation is significantly higher: 0.19 for Facebook posts against 0.07 for Silvia Romano tweets.

The qualitative analysis was carried out by manually inspecting the highlighted chunks from couples of annotations that scored particularly high or particularly low on the measure of similarity. From the quantitative analysis, it emerges that annotators obtained a higher agreement selecting span from Facebook posts than from tweets about Silvia Romano: such a result could be explained by the different domains of the Silvia Romano dataset. In fact, even though the tweets mention Silvia Romano, this dataset also contains many offensive comments and words on Islamophobia

and choices made by the Italian government, and not always as offensive comments against Silvia Romano.

Looking at annotations from this last dataset, the comments with more overlap are often those in which the highlighted spans coincide with the entire text. Moreover, it is possible to observe that some of these comments are directed to Silvia Romano, specifically on her body (traces of body-shaming are evident), others show scepticism about Stockholm syndrome, and some are explicit death threats (see table 2 Silvia Romano Id 1040). On the other hand, the comments with less overlap are often those pertaining different domains, such as the government, or religion, which were not the main target of the annotation task (see table 2 Silvia Romano Id 395).

Source	Text	Chunk 1	Chunk 2
Silvia Romano Id 1040	Silvia Romano stai attenta che se si dovesse manifestarsi qualche attentato da parte del gruppo in cui ti sei convertita,ti troveremo e ti faremo a pezzi,altro che sciabole...	*Silvia Romano stai attenta che se si dovesse manifestarsi qualche attentato da parte del gruppo in cui ti sei convertita,ti troveremo e ti faremo a pezzi,altro che sciabole...*	*Silvia Romano stai attenta che se si dovesse manifestarsi qualche attentato da parte del gruppo in cui ti sei convertita,ti troveremo e ti faremo a pezzi,altro che sciabole*
Silvia Romano Id 395	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.*

Table 2: Example of comments with more and less agreement for Silvia Romano dataset.

Regarding Facebook dataset, the comments with more agreement are generally shorter, so again the annotators selected chunks corresponding to the full phrases, it is also noteworthy that almost all of the comments with a very high degree of similarity refer to physical aspects (see table 3 Facebook Id 299). While the comments with less overlap seem to be longer and generally with more offensive terms (see table 3 Facebook Id 77).

Source	Text	Chunk 1	Chunk 2
Facebook Id 299	Bruttissima fa schifo il suo viso sembra plastica ☠️	*Bruttissima fa schifo il suo viso sembra plastica*	*Bruttissima fa schifo il suo viso sembra plastica*
Facebook Id 77	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!!*

Table 3: Example of comments with more and less agreement for Facebook dataset.

6. Conclusion and future work

In this work we present two corpora developed through an experimental annotation task designed to explore disagreement among annotators. For a given comment, the annotation procedure consisted in selecting one or more chunks from each text that is regarded as misogynistic and establishing whether a gender stereotype is present. As a result of the annotation task, 2,207 annotations of tweets about Silvia Romano and 4,942 annotations of Facebook posts were collected.

The analysis of annotations showed a high level of disagreement in both tasks. From the quantitative analysis it emerged that annotators obtained a higher agreement when selecting span from Facebook posts than from tweets about Silvia Romano: such a result could be explained by the different domains of the Silvia Romano dataset. In fact, even though the tweets mention Silvia Romano, this dataset also contains many offensive comments and words on Islamophobia and choices made by the Italian government, and not always as offensive comments against Silvia Romano. In general, the comments with more overlap are often those in which the highlighted spans coincide with the entire text, while the comments with less overlap tend to be longer and generally contain more offensive terms.

Future work will focus on expanding this work into different domains, in order to better analyse how disagreement impacts on computational resources and try to integrate disagreement into modelling and evaluation.

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.

Beigman Klebanov B., Beigman E., and Diermeier D. 2008. Analyzing disagreements. In Coling 2008: *Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 2–7, Manchester, UK. Coling 2008 Organizing Committee.

Bowman, S. R., & Dahl, G. E. (2021). What Will it Take to Fix Benchmarking in Natural Language Understanding?. arXiv preprint arXiv:2104.02145.

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.

Landis JR., Koch GG. 1977. The measurement of observer agreement for categorical data. Biometrics. 1977 Mar;33(1):159-74. PMID: 843571.

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.

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

Pavlopoulos, J., Sorensen, J., Laugier, L., & Androutsopoulos, I. (2021, August). Semeval-2021 task 5: Toxic spans detection. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021) (pp. 59-69).

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.

Tontodimamma A., Fontanella L., Anzani S., Basile V. (2022). An Italian lexical resource for incivility detection in online discourses. Quality & Quantity. https://doi.org/10.1007/s11135-022-01494-7.