An Experimental Annotation Task Investigating Annotator Agreement within a Misogynistic Dictionary and Corpus

Alice Tontodimamma, Elisa Ignazzi, Stefano Anzani, Lara Fontanella, Simone Di Zio

G. d'Annunzio University, Chieti-Pescara, Italy

IES2023
30 August -1 September, 2023

Outline

Data Labelling

- **Data labelling**: process of assigning meaningful annotations or tags to raw data, enhancing its clarity and usefulness for various applications, including medicine and psychology, but also research in content analysis and corpus linguistics.

Within the realm of NLP and AI, the emphasis has shifted to Machine Learning, making the development of datasets for training and assessing AI systems a pivotal undertaking.

- AI systems **learn patterns from data**.
- AI success hinges on the availability of **high-quality labelled datasets** for training and evaluation.
- **Accurate data labelling** is vital for **meaningful insights and reliable AI models**.

Data Labelling and Inter-rater Agreement

- **Inter-rater agreement measures quantify the level of consensus and the consistency between multiple annotators**.
- Inter-rater agreement is a critical quality metric for data labelling:
  - High inter-rater agreement indicates consistent and reliable annotations;
  - Low inter-rater agreement raises concerns about ambiguity and reliability of the annotation process.
- ***Assumption***: if different coders consistently generate comparable outcomes, we can deduce that they have internalised a similar understanding of the annotation guidelines.

Observed agreement

Given the item set $\{i \in I\}$ to be annotated into $\{k \in K\}$ categories and the set of coders $\{c \in C\}$, the observed agreement is computed as

$$A_o = \frac{1}{|I|} \sum_{i \in I} agr_i$$

2-coders

$$agr_i = \begin{cases} 1 & \text{the 2 coders assign the same category } k \text{ to item } i \\ 0 & \text{the 2 coders assign different categories to item } i \end{cases}$$

multi-coders

$$agr_i = \frac{1}{\binom{|C|}{2}} \sum_{k \in K} \binom{n_{ik}}{2} = \frac{1}{|C|(|C|-1)} \sum_{k \in K} n_{ik}(n_{ik}-1)$$

- n_{ik} =number of times item i is classified into category k

Observed agreement: limitations

Observed agreement enters in the computation of all the measures of agreement.

By itself, observed agreement does not provide values suitable for cross-study comparison, since **some agreement is due to chance**.

The extent of this random agreement is influenced by two factors that differ between studies [Paun et al., 2022]:

- percentage agreement is biased in favour of dimensions with a small number of categories;
- percentage agreement does not correct for the distribution of items among categories: we expect a higher percentage agreement when one category is much more common than the other.

In order to get figures that are comparable across studies, **observed agreement has to be adjusted for chance agreement.**

Chance-corrected inter-rater agreement coefficients

$$\phi = \frac{A_o - A_e}{1 - A_e}$$

- A_e: expected agreement: amount of agreement we would expect to see if the coders were making arbitrary label choices
- $A_o - A_e$: measures how much agreement beyond chance was actually found
- $1 - A_e$: measures how much agreement over and above chance is attainable
- ϕ: measures which proportion of the possible agreement beyond chance was actually observed

Chance agreement: requires a model that specifies the notion of arbitrary agreement and each coefficient specifies this notion differently.

Coefficients of agreement for computational linguistics tasks

- AIM of agreement measures in **computational linguistics tasks**: to infer about the **reliability of large-scale annotation processes**.

Most appropriate inter-rater agreement measures: *shared-distributions coefficients*

- Fleiss' κ
- Krippendorff's α.

For binary annotation schemes, the two indexes give similar results so we focus on Fleiss' κ *.*

Limitation:

- **Prevalence problem**: if a disproportionate amount of the data falls under one category (skewed distribution), then the expected agreement is very high, so in order to demonstrate high reliability an even higher observed agreement is needed [Di Eugenio, 2000, Di Eugenio and Glass, 2004].

Alternative: Probabilistic models of agreement

All classical coefficients of agreement estimate expected agreement on the entire set of items.

Probabilistic models distinguish between:

- **Easy items** in which deliberate consensus among the annotators can be observed;
- **Difficult items** in which the annotations present disagreement or there is random agreement.

Gwet AC_1 [Gwet, 2008] estimates A_e for the difficult items only: easy items may be disregarded on the grounds that any agreement will not be by chance.

Fleiss' κ [Fleiss, 1971]: expected agreement

Fleiss' κ coefficient is a generalisation of Scott's π [Scott, 1955] defined for 2 coders.

2-coders

$$A_e^{(S)} = \sum_{k \in K} P(k|c_1)P(k|c_2)$$

- $P(k|c_1) = P(k|c_2) = \widehat{P(k)} = \frac{n_k}{2|I|}$
- n_k =total number of assignments to k by both coders

multi-coders

$$A_e^{(F)} = \sum_{k \in K} \left(\widehat{P(k)}\right)^2$$

- $\widehat{P(k)} = \frac{n_k}{|I||c|}$
- n_k =total number of assignments to k by all the coders

Gwet's AC_1 [Gwet, 2008]: expected agreement

$|C| = 2$ raters classify $|I|$ items into either category '0' or '1' independently

- G= {The two raters c_1 and c_2 agree}
- R= {A rater (c_1, or c_2) or both performs a random rating}

$$P(G|R) = 2 * 0.5^2 = 0.5; \quad P(R) \approx \frac{\pi_1(1-\pi_1)}{0.5(1-0.5)} = 4\pi_1(1-\pi_1)$$

$$A_e^{(G)} = P(G \cap R) = P(G|R)P(R) \approx 2\pi_1(1-\pi_1)$$

The probability π_1 can be estimated from sample data as

$$\hat{\pi}_1 = 0.5(p_{c_1,1} + p_{c_2,1})$$

$$p_{c_1,1} = \frac{n_{c_1,1}}{|I|}; \quad p_{c_2,1} = \frac{n_{c_2,1}}{|I|}$$

Gwet extends the AC_1 coefficient also to multiple raters.

Simulation studies

In order to understand the effect of the distribution of the ratings among the categories on the inter-agreement coefficients we implemented two simulation studies considering

- a binary annotation scheme ($K = 2$)
- $I = 100$ items to be annotated
- 2-raters and multi-raters scenarios

Fleiss'κ and Gwet's AC_1 were computed using the irrAC package [Gwet, 2019] for the R environment.

Simulation study for 2-raters

To simulate the annotations for 2 raters on $I = 100$ items, we

1. simulate two underlying zero-mean normal variables with a given correlation matrix $\mathbf{\Sigma}$ where $E(Z_1, Z_2) = \rho$

$$\boldsymbol{Z} \sim \mathcal{N}_2(\mathbf{0}, \mathbf{\Sigma})$$

2. obtain two binary annotation variables considering the threshold model

$$x_{c,i} = \begin{cases} 0 & \text{if} \quad z_{c,i} \leq \gamma \\ 1 & \text{otherwise} \end{cases}$$

for $c = 1, 2$ and $i = 1, \ldots, I$.

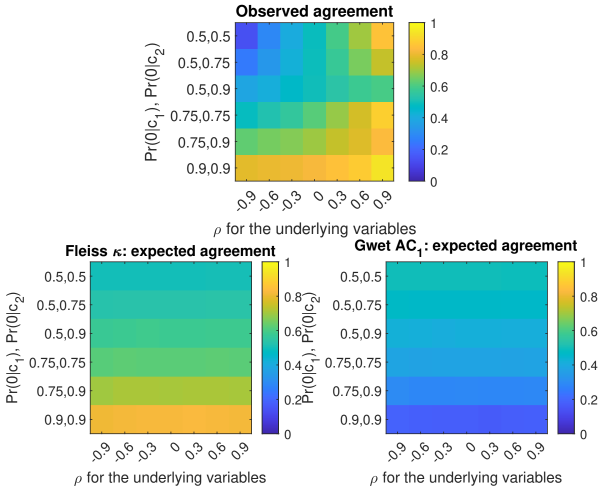
Simulation parameters

- $\rho = (-0.9, -0.6, -0.3, 0, 0.3, 0.6, 0.9)$
- $\gamma = (\Phi^{-1}(0.50), \Phi^{-1}(0.75), \Phi^{-1}(0.90))$

Monte Carlo repetitions for each combination of parameters: 100

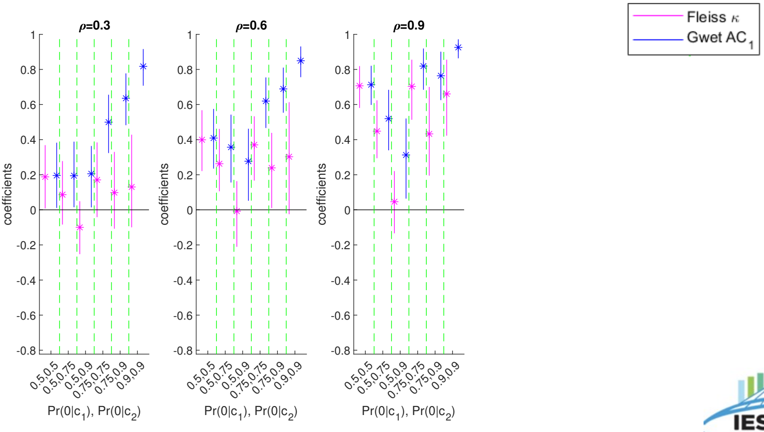
Observed A_o and expected A_e agreement for Fleiss' κ and Gwet's AC_1

mean values across the 100 repetitions



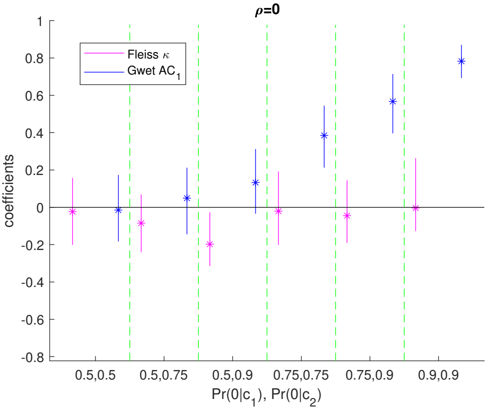
Positive correlated underlying variables:

Coefficient values and 95% confidence intervals for Fleiss' κ and Gwet's AC_1



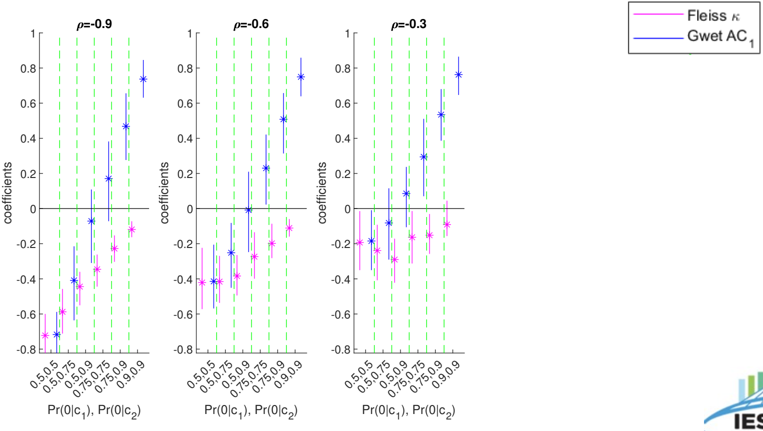
Independent underlying variables:

Coefficient values and 95% confidence intervals for Fleiss' κ and Gwet's AC_1



Negative correlated underlying variables:

Coefficient values and 95% confidence intervals for Fleiss' κ and Gwet's AC_1



Simulation study for multiple raters

To simulate the annotations for N_c raters on $I = 100$ items with a binary annotation scheme, we simulate a $I \times 2$ table where for each item we record the number of raters for each category.
For $i = 1, \ldots, I$

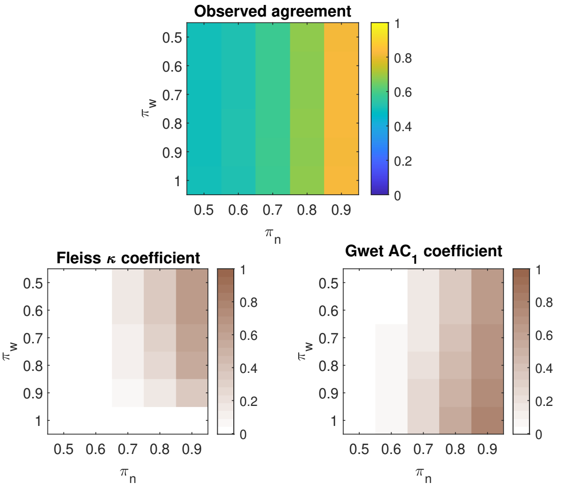
1. simulate
$$n_i \sim Binomial\,(N_c, \pi_n)$$
2. simulate
$$w_i \sim Bernoulli\,(\pi_w)$$
3. set the number of annotations for the two categories as
$$n_{i,0} = w_i \cdot n_i + (1 - w_i) \cdot (N_c - n_i)$$
$$n_{i,1} = N_c - n_{i,0}$$

Simulation parameters

- $N_c = 10$
- $\pi_n = (0.5, 0.6, 0.7, 0.8, 0.9)$
- $\pi_w = (0.5, 0.6, 0.7, 0.8, 0.9, 1)$

Observed A_o and coefficient values for Fleiss' κ and Gwet's AC_1

mean values across the 100 repetitions



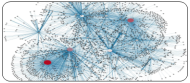
ICOMIC Project

ICOMIC: Identifying and Countering Online Misogyny

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

Funded by the European Union

Misogynistic speech detection



Identification of producers of misogynistic content

Countering misogynistic speech

Project objectives and annotation tasks

Misogynistic speech detection

Aim

- Build a new **lexical resource** specifically designed to include terms expressing hatred towards women
- Create a **corpus** of comments shared on Twitter, Facebook, Instagram, YouTube, and Reddit **annotated for misogyny**.
- Compare the performance of **interpretable machine learning model** (e.g., naive Bayes), exploiting lexical, lexicon-based, and sentiment analysis features, with xAI approaches.

Annotation tasks

Task 1: Annotating online comments for misogyny: *2-raters*

- 2 rounds of annotation
- each round: 3000 comments extracted from Twitter, Facebook and Instagram
- 10 trainees divided into 5 groups. Each trainee pair annotated independently 300 comments in each round.
- Binary annotation: every annotator was asked to annotate each comment for misogynistic content

Task 2: Annotating lexicon for misogyny: *multi-raters*

- 1200 terms exctracted from the Revised Hurtlex lexicon [Tontodimamma et al., 2023]
- 6 trainees
- Binary annotation: every annotator was asked to annotate each term for misogynistic content
- If the term was misogynistic, annotators were told to choose a subcategory of misogyny

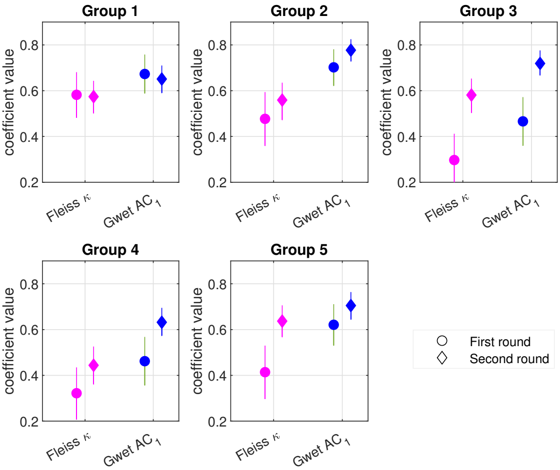
Annotating online comments for misogyny: first round

group	$p(X_{c1}=1)$	$p(X_{c2}=1)$	$p(X=1)$	A_o		A_e	coefficient
1	0.377	0.273	0.325	0.817	Fleiss' κ	0.561	0.582
					Gwet's AC_1	0.439	0.673
2	0.177	0.300	0.238	0.810	Fleiss' κ	0.637	0.477
					Gwet's AC_1	0.363	0.702
3	0.197	0.433	0.315	0.697	Fleiss' κ	0.568	0.297
					Gwet's AC_1	0.432	0.466
4	0.197	0.463	0.330	0.700	Fleiss' κ	0.558	0.322
					Gwet's AC_1	0.442	0.462
5	0.173	0.363	0.268	0.770	Fleiss' κ	0.607	0.414
					Gwet's AC_1	0.393	0.621

Annotating online comments for misogyny: second round

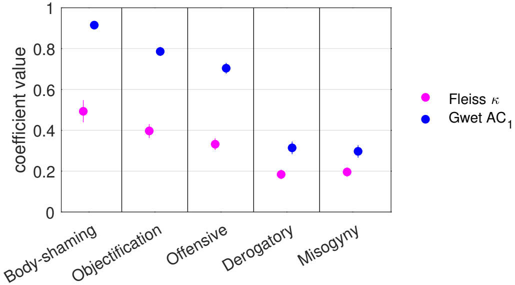
group	$p(X_{c1}=1)$	$p(X_{c2}=1)$	$p(X=1)$	A_o		A_e	coefficient
1	0.327	0.358	0.343	0.808	Fleiss' κ	0.550	0.574
					Gwet's AC_1	0.450	0.651
2	0.238	0.188	0.213	0.852	Fleiss' κ	0.664	0.559
					Gwet's AC_1	0.336	0.777
3	0.278	0.280	0.279	0.832	Fleiss' κ	0.598	0.581
					Gwet's AC_1	0.402	0.719
4	0.180	0.368	0.274	0.779	Fleiss' κ	0.602	0.444
					Gwet's AC_1	0.398	0.632
5	0.335	0.345	0.340	0.837	Fleiss' κ	0.551	0.637
					Gwet's AC_1	0.449	0.705

Training effect on inter-rater agreement



Annotating lexicon for misogyny

Task	p(X=1)	A_o	Fleiss; κ A_e	Fleiss; κ Coefficient	Gwet's AC_1 A_e	Gwet's AC_1 Coefficient
Body-shaming	0.08	0.93	0.86	0.49	0.14	0.92
Objectification	0.16	0.84	0.74	0.40	0.26	0.79
Offensive	0.19	0.80	0.69	0.33	0.31	0.70
Derogatory	0.35	0.63	0.54	0.18	0.46	0.31
Misogyny	0.63	0.63	0.53	0.20	0.47	0.30



Conclusions

- If a disproportionate amount of the data falls under one category (skewed distribution), then the expected agreement computed through classical coefficients of agreement is very high, so in this case, it is better to use Gwet's AC_1 coefficient.
- The distributions of the first and second annotation tasks are skewed.
  - First annotation task: positive training effect on both inter- rater agreement measures.
  - Second annotation task: some categories show coefficient values higher because are very likely easier to annotate, such as the body-shaming category.
- One of the main goals of the ICOMIC project is to release reliable data, for that reason, it was useful to explore the behaviour inter-rater agreement measures to choose the most suitable according to the annotation task, category numbers and their distributions.

References

B. Di Eugenio. On the Usage of Kappa to Evaluate Agreement on Coding Tasks. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, Athens, Greece, 2000. European Language Resources Association (ELRA).

B. Di Eugenio and M. Glass. quibs and Discussions: The Kappa Statistic: A Second Look. *Computational Linguistics*, 30(1):95–101, 2004.

J.L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

K. L. Gwet. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48, 2008.

K. L. Gwet. *irrCAC: Computing Chance-Corrected Agreement Coefficients (CAC)*, 2019. URL `https://CRAN.R-project.org/package=irrCAC`. R package version 1.0.

S. Paun, R. Artstein, and M. Poesio. *Statistical Methods for Annotation Analysis*. Springer, 2022.

W. A. Scott. Reliability of Content Analysis:The Case of Nominal Scale Coding. *Public Opinion Quarterly*, 19(3):321–325, 1955.

A. Tontodimamma, L. Fontanella, S. Anzani, and V. Basile. An Italian lexical resource for incivility detection in online discourses. *Quality and Quantity*, 57:3019–3037, 2023. doi: 10.1007/s11135-022-01494-7.