Manual Label Evaluation: Comparing Labeling Rounds (Interrater/Intercoder Reliability)

Between two individual Coders

```
Cohen's Kappa          0.68323

Percent Agreement    89.2157

Equal Labels Count   273

Total Labels         306
```

- Cohen's Kappa: Represents the amount of agreement that can be expected from random chance, and 1 represents perfect agreement between the raters. $>0.6 =$ **moderate agreement**
- Percent Agreement: The percentage of times the two coders agreed on the label.

Category Specific Agreement

From the agreement package: "Specific agreement is an index of the reliability of categorical measurements. It describes the amount of agreement observed with regard to each possible category. With two raters, the interpretation of specific agreement for any category is the probability of one rater assigning an item to that category given that the other rater has also assigned that item to that category. With more than two raters, the interpretation becomes the probability of a randomly chosen rater assigning an item to that category given that another randomly chosen rater has also assigned that item to that category. When applied to binary (i.e., dichotomous) data, specific agreement on the positive category is often referred to as positive agreement (PA) and specific agreement on the negative category is often referred to as negative agreement (NA). In this case, PA is equal to the F1 score frequently used in computer science." "

```
Call:
cat_specific(.data = data)

Objects =   306
Raters =    2
Categories =    {0, 1}

Category-Specific Agreement with Bootstrapped CIs

     Estimate   lower   upper
0       0.931   0.906   0.953
1       0.752   0.667   0.830
```

Between two individual coders and a labeling version that multiple coders worked on in intervals

```
+-------------------+-----------+
| Metric            |     Value |
+===================+===========+
| Fleiss Kappa      |  0.681166 |
+-------------------+-----------+
| Percent Agreement | 88.4526   |
+-------------------+-----------+
```

- Fleiss' Kappa: calculates the degree of agreement in classification over that which would be expected by chance. $>0.6 =$ **Moderate Agreement**

Category Specific Agreement

```
Call:
cat_specific(.data = data)

Objects =   306
Raters =    3
Categories =    {0, 1}

Category-Specific Agreement with Bootstrapped CIs
```

```
     Estimate   lower   upper
0       0.924   0.904   0.943
1       0.757   0.689   0.817
```

Conclusion

According to McHugh (2012): "Perhaps the best advice for researchers is to calculate both percent agreement and kappa. If there is likely to be much guessing among the raters, it may make sense to use the kappa statistic, but if raters are well trained and little guessing is likely to exist, the researcher may safely rely on percent agreement to determine interrater reliability."

AWO coders were trained and did not guess, so we rely on percent agreement. Therefore, interrater/intercoder reliability is sufficient with strong agreement overall. For the specific labels, there was strong agreement for the negative label 0 and moderate agreement for the positive label.

Rows where the two trained AWO coders both differed from the original labeling round

```
Rows where the two independent coders both differed from the original labeling round: 20


title                                                           coder_3   coder_1_2
--------------------------------------------------------------- --------- -----------
Förderung von Integrationsagenturen für die Belange von Mens        1            0
Kurse zur Alphabetisierung für Asylsuchende                         1            0
Zuwendungen im Rahmen des Asyl-, Migrations- und Integration        1            0
Soziale Beratung von Flüchtlingen in Nordrhein-Westfalen            1            0
Förderung der Durchführung von Asylverfahrensberatung (AVB)         0            1
Maßnahmen und Projekte gemäß § 96 des Bundesvertriebenengese        1            0
Förderungen im Rahmen des Europäischen Sozialfonds Plus (ESF        1            0
Förderung von Forschungsprojekten zum Thema „Islamismus: Aus        0            1
Erstorientierungskurse für Schutzsuchende und Zugewanderte          1            0
Berufliche Wiedereingliederung von Inhaftierten und Haftentl        1            0
Demenz und Teilhabe (DEMTeil)                                       1            0
Netzwerke integrieren Geflüchtete in den regionalen Arbeitsm        1            0
ESF-Bundesprogramm - „rückenwind3 für Vielfalt, Wandel und Z        1            0
Sportförderrichtlinie                                               1            0
Teilhabe am gesellschaftlichen Leben durch Vermittlung von I        1            0
Justiz (2021-2027)                                                  1            0
```

```
Europäischer Sozialfonds Plus (ESF+) in Rheinland-Pfalz          1          0
Europäischer Sozialfonds (ESF+) (2021-2027)                      1          0
Bundesaltenplan                                                   1          0
Förderung der Teilhabe von Menschen mit Beeinträchtigungen d     1          0
```

Original corrected label round as ground truth

```
+-----------+-----------+-----------+-----------+
| Metric    |   Coder 1 |   Coder 2 |   Average |
+===========+===========+===========+===========+
| Accuracy  |  0.882353 |  0.879085 |  0.880719 |
+-----------+-----------+-----------+-----------+
| F1 Score  |  0.756757 |  0.76129  |  0.759024 |
+-----------+-----------+-----------+-----------+
| Recall    |  0.888889 |  0.842857 |  0.865873 |
+-----------+-----------+-----------+-----------+
| Precision |  0.658824 |  0.694118 |  0.676471 |
+-----------+-----------+-----------+-----------+
```