INTERNATIONAL DESIGN CONFERENCE – DESIGN 2020
https://doi.org/10.1017/dsd.2020.153

INVESTIGATING INTER-RATER RELIABILITY OF QUALITATIVE TEXT ANNOTATIONS IN MACHINE LEARNING DATASETS

N. El Dehaibi ✉ and E. F. MacDonald

Stanford University, United States of America

✉ ndehaibi@stanford.edu

Abstract

An important step when designers use machine learning models is annotating user generated content. In this study we investigate inter-rater reliability measures of qualitative annotations for supervised learning. We work with previously annotated product reviews from Amazon where phrases related to sustainability are highlighted. We measure inter-rater reliability of the annotations using four variations of Krippendorff's U-alpha. Based on the results we propose suggestions to designers on measuring reliability of qualitative annotations for machine learning datasets.

Keywords: artificial intelligence (AI), big data analysis, qualitative annotations, design methods

1. Introduction

The rapid growth in online user generated content and advancements in machine learning algorithms have enabled new approaches for designers to learn about customer needs. Designers traditionally conduct interviews, surveys, focus groups, or simply observe customers in a target context to better understand their needs. Designers are also now able to identify important customer insights from sources like product reviews or tweets using machine learning models and natural language processing techniques. These approaches are potentially faster, more cost-effective, and address some biases compared to traditional approaches like surveys or interviews, but also introduce new challenges (Tuarob and Tucker, 2015).

In supervised learning designers provide samples of input and output data to build a model. For example, Stone and Choi annotate tweets about phone products based on positive, negative, and neutral emotions in the tweets (Stone and Choi, 2013). In this example the inputs are the tweets and the outputs are the annotations. A common challenge with this type of dataset is measuring the reliability of the annotations since the quality of the model depends on it. In machine learning research, a common way to evaluate the dataset is by looking at the evaluation metrics of the model such as precision, recall, F1 (Jurafsky and Martin, 2017). While these metrics can provide external validity for a model, they are not commonly used in the design research space.

A more common approach for assessing reliability of annotator data in design research is inter-rater reliability (IRR) which provides an internal validity check. With IRR the responses from different annotators are compared using statistical analyses (Gwet, 2014). For example, Toh et al. use IRR to measure the agreement between two annotators rating a set of electric toothbrush design concepts (Toh et al., 2014). By achieving high IRR measures, the authors can have confidence in their research approach and results collected. For an overview of IRR, please refer to section 2.

In this paper we explore IRR as a measure of reliability of qualitative annotations in machine learning datasets. The goal is to provide designers with familiar metrics besides machine learning metrics to assess reliability of annotator data. We work with previously collected annotations of product reviews from Amazon where phrases relevant to sustainability are identified and highlighted (El Dehaibi et al., 2019). This is a highly qualitative annotation task because sustainability is a complex and often subjective concept. We use IRR to measure the degree of agreement among annotators and discuss the results given our context. The rest of the paper is organized as follows: in section 2 we provide an overview on IRR, in section 3 we describe our research approach, in section 4 we present the results, we discuss the results in section 5, and we conclude the paper in section 6.

2. Overview of inter-rater reliability measures

Inter-rater reliability (IRR) is a statistical measure of the degree of similarity between the results of different raters' (in this case, annotators), judging tasks. These tasks may involve sorting, judging on a scale, and parsing phrases. Based on these different tasks, raters may also be known as "coders", "judges", "observers", or "annotators". The IRR scale ranges from below 0 (denoting no agreement) up to 1 (denoting perfect agreement). The idea behind IRR is that the more agreement there is between the raters, the higher the confidence we can have in what the raters provide. Several measures exist for IRR, the simplest being a joint probability agreement which measures the percentage of observed agreement. Most IRR measures correct for expected agreement by chance and are considered to be a more robust estimate of the agreement, otherwise the agreement measure is overestimated (Hallgren, 2012). We discuss some of the commonly used IRR measures below.

2.1. Cohen's kappa

Cohen's kappa measures the IRR between two raters for categorical items (Cohen, 1960). Recent examples of research using Cohen's kappa include studying reliability of coders assigning categories to audio files (Kennedy et al., 2019), or categorizing photos based on visitor behavior in public parks (Liang et al., 2019). Cohen's kappa is a function of p_o, the relative observed agreement, and p_e, the expected hypothetical agreement by chance, as shown below (Equation 1).

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{1}$$

The observed agreement is calculated using Equation (2),

$$p_0 = \frac{count\ of\ agreements}{count\ of\ agreements + count\ of\ disagreements} \tag{2}$$

and the expected agreement by chance is calculated using Equation (3),

$$p_e = \frac{1}{N^2} \sum_k n_{k_1} n_{k_2} \tag{3}$$

where k is the number of categories, N is the number of items, n_{k_1} is the number of times rater 1 selected category k, and n_{k_2} is the number of times rater 2 selected category k. The advantage of Cohen's kappa is that it corrects for the expected agreement by chance and is therefore a robust estimate, but the disadvantage is that it is limited to only two raters for categorical items.

2.2. Fleiss' kappa

Fleiss' kappa extends Cohen's kappa to work with any fixed number of raters for categorical items (Fleiss, 1971). Recent examples of research using Fleiss' kappa include studying how well participants can read facial expressions (Rash et al., 2019), and evaluating psychometric perceptions of satisfaction questionnaires for patients and family members (Lai et al., 2019). Fleiss' kappa takes the same form as Equation (1), however, the observed and expected agreements are calculated differently as shown in Equations (4) and (5), respectively.

$$p_0 = \frac{1}{Nn(n-1)} \left(\sum_{i=1}^{N} \sum_{j=1}^{k} n_{ij}^2 - Nn \right) \tag{4}$$

where N is the number of raters, n is number of items, k is the number of categories, i is the index for each rater, and j is the index for each category.

$$p_e = \sum_{j=1}^{k} \left(\frac{1}{Nn} \sum_{i=1}^{N} n_{ij}\right)^2 \qquad (5)$$

The advantage of Fleiss' kappa is that it is not limited to only two raters, but the disadvantage is that it can only be used to measure reliability of categorical items.

2.3. Krippendorff's U-alpha

Krippendorff's U-alpha measures the IRR for any number of raters and different types of data including both nominal and ordinal (Krippendorff, 2004). It is commonly used for measuring reliability of qualitative text analysis data such as highlighted text. Recent examples of works that have used Krippendorff's U-alpha include identifying arguments in portions of text (Stab and Gurevych, 2014), and identifying policy issues in news articles (Card et al., 2015).

For a given text of length L, Krippendorff's U-alpha quantifies highlighted text by measuring where a highlight starts, *b*, and how long the highlight is, *l*, for each category (see Figure 1).

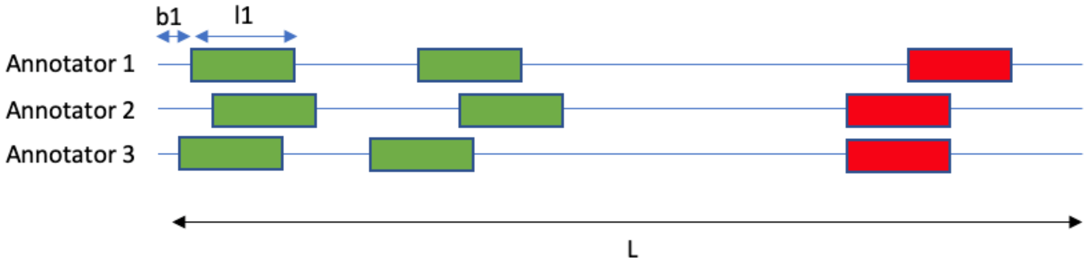

Figure 1. Quantifying text annotations for Krippendorff's U-alpha

In Figure 1 there are three annotators for a text of length L and two categories (red and green). The highlights are quantified in terms of *b* and *l* and the differences between annotators is measured to calculate agreement For a given category c, Krippendorff's U-alpha is calculated using the observed disagreement, D_{oc}, and expected disagreement, D_{ec}, as shown in Equation (6).

$$\alpha_c = 1 - \frac{D_{oc}}{D_{ec}} \qquad (6)$$

For a given category c, D_{oc} is calculated as shown in Equation (7),

$$D_{oc} = \frac{\sum_{i=1}^{m} \sum_g \sum_{j=1|j\neq i}^{m} \sum_h \delta_{cigjh}^2}{m(m-1)L^2} \qquad (7)$$

δ_{cigjh}^2 is the squared difference between annotation g and annotation h corresponding to any two observers i and j, respectively, m is the number of raters, and L is the length of the given data. Length L and difference δ_{cigjh} is typically measured using letter counts as a unit of length. The difference δ_{cigjh} is calculated in Equation (8) as follows:

$$\delta_{cigjh} = \left\{\begin{matrix} \left(b_{cig} - b_{cjh}\right)^2 + \left(b_{cig} + l_{cig} - b_{cjh} - l_{cjh}\right)^2 \; iff \; v_{cig} = v_{cjh} = 1 \; and - l_{cig} < b_{cig} - b_{cjh} < l_{cjh} \\ l_{cig}^2 \; iff \; v_{cig} = 1, v_{cjh} = 0 \; and \; l_{cjh} - l_{cig} \geq b_{cig} - b_{cjh} \geq 0 \\ l_{cjh}^2 \; iff \; v_{cig} = 0, v_{cjh} = 1 \; and \; l_{cjh} - l_{cig} \leq b_{cig} - b_{cjh} \leq 0 \\ 0 \; otherwise \end{matrix}\right\} \qquad (8)$$

where b denotes the beginning of a highlight for a given rater, l is the length of a given highlight, and v is binary denoting if a section is a highlight ($v = 1$) or not a highlight ($v = 0$). For text data, b and l are defined in terms of letter counts. The expected disagreement is then defined as shown in Equation (9),

$$D_{ec} = \frac{\frac{2}{L} \sum_{i=1}^{m} \sum_g v_{cig} \left[\frac{N_c - 1}{3}\left(2l_{cig}^3 - 3l_{cig}^2 + l_{cig}\right) + l_{cig}^2 \sum_{j=1}^{m} \sum_h \left(1 - v_{cjh}\right)\left(l_{cjh} - l_{cig} + 1\right) \; iff \; l_{cjh} \geq l_{cig}\right]}{mL(mL-1) - \sum_{i=1}^{m} \sum_g v_{cig} l_{cig}\left(l_{cig} - 1\right)} \qquad (9)$$

where L is the total length of the text (letter counts). To calculate Krippendorff's U-alpha for multiple categories, the observed and expected disagreements for each category are summed as shown in Equation (10).

$$\alpha = 1 - \frac{\sum_c D_{oc}}{\sum_c D_{ec}} \tag{10}$$

Based on the above explanations of different IRR metrics, we choose to focus on Krippendorff's U-alpha because it is the most generalizable approach for different types of data and enables us to calculate reliability of qualitative highlighted text. Krippendorff's U-alpha was created to measure agreement between raters for qualitative text, but it is unclear if a high degree of agreement is desirable in the context of machine learning datasets. In this paper we investigate the implications of Krippendorff's U-alpha measure when annotating text for machine learning datasets.

3. Research approach

In this study we use annotations of product reviews of French Press coffee makers collected in (El Dehaibi et al., 2019). Annotators were recruited from Amazon Mechanical Turk and participated in a Qualtrics survey where they were briefly trained on either social, environmental, and economic sustainability. There were three versions of the survey to account for each sustainability aspect. After completing the training, annotators were asked to highlight phrases related to a sustainability aspect in product reviews and to rate the positive and negative emotions in the phrases they highlighted. The authors adhered to common practice for high quality responses from Amazon Mechanical Turk as outlined by Paolacci and Chandler (Paolacci and Chandler, 2014) and Goodman and Paolacci (Goodman and Paolacci, 2017). Bonus compensation was also offered for annotators to incentivize high quality work. Some of the reviews where annotated by two or more participants: social (449 reviews), environmental (404 reviews), and economic (436 reviews), for a total of 1289 reviews that we use in this IRR study. Note that for these 1289 reviews, there are two to three annotations per review.

We calculate IRR measures on the annotated reviews using Krippendorff's U-alpha as defined in Equations (7-10). The annotations are split into two categories, positive emotion and negative emotion. As a baseline we use letter counts in Equation (8) to measure differences between annotations and calculate Krippendorff's U-alpha. In addition to the baseline we implement some variations so that we may tune how sensitive the IRR measure is to differences between annotators. The baseline measure and variations implemented are explained below.

- **Baseline (Letter counts):** We calculate Krippendorff's U-alpha using letter counts as unit of difference measure between annotations (the smallest unit of length). The difference between annotator highlights is counted by letters.
- **Letter counts with natural language processing (NLP)**: Similar to the baseline, we use letter counts to measure length and differences between annotations, however we first pre-process the reviews with natural language processing. This includes lowercasing, removing white spaces, numbers, punctuation, and stop words, lemmatizing, and stemming the words in the reviews. The intuition of using NLP is that it can remove potential noise in the annotations.
- **Word counts:** We calculate Krippendorff's U-alpha using word count to measure length and differences. Although Krippendorff's alpha adjusts based on length of the overall text, the intuition of using word counts is that it may make the overall calculation less sensitive to distances between annotator highlights.
- **Word counts with NLP:** We calculate Krippendorff's U-alpha using word count to measure length and differences, and also pre-process the reviews with natural language processing. We implement the same NLP steps as in "Letter counts with NLP" but using word counts instead. We intuit that NLP may have a bigger impact when looking at word counts and better allow us to tune the outputs as needed.

For this study we calculate Krippendorff's U-alpha for three sets of 400 to 450 reviews (a set for each sustainability aspect) using the above four measures. We developed a Python code to calculate

Krippendorff's U-alpha measures on the annotations, available on GitHub[1]. We compare the different IRR variations to determine if we can tune the output to provide insight on the reliability of the annotator data. Based on the results, we then discuss these measures in the context of qualitative annotations for machine learning datasets.

4. Results

The mean IRR scores from 400 to 450 reviews for each sustainability aspect are shown in Figure 2 below. Along the horizontal axis we have three sets of horizontal bars, one for each sustainability aspect. Within each set are the results of the different variations of Krippendorff's U-alpha variations described in section 3.

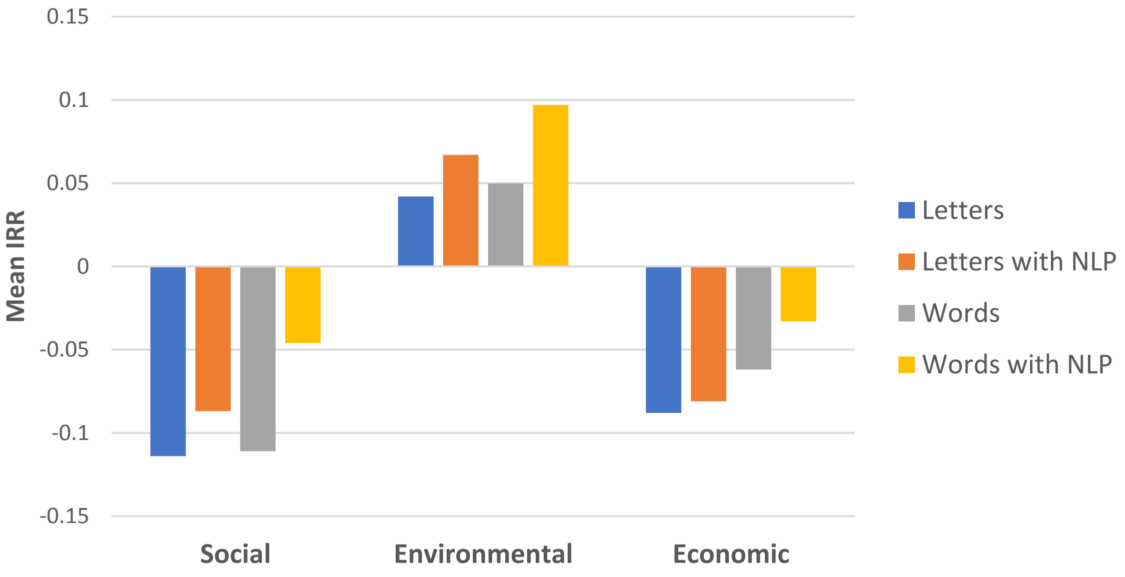

Figure 2. Mean IRR scores for each sustainability aspect

In Figure 2 we see the IRR scores for environmental sustainability in the middle set were highest on average, ranging from 0.042 to 0.097. The IRR scores for economic sustainability in the right set were second highest on average, ranging from -0.088 to -0.033. The IRR scores for social sustainability in the left set were the lowest on average, ranging from -0.114 to -0.046. We also see that pre-processing the reviews with natural language processing and looking at word counts resulted in the highest IRR scores on average. Pre-processing reviews with NLP and looking at letter counts also increased scores, but not as much.

The difference in Krippendorff's U-alpha between word counts compared to letter counts is negligible in the absence of NLP. For example, the mean IRR scores for environmental sustainability are 0.042 and 0.0498 for letter counts and words counts respectively. The negligible difference without NLP was expected because despite having smaller distances with word counts, the overall difference gets normalized by a smaller review length compared to when looking at letter counts.

We also see in Figure 2 that on average the IRR scores revolve around 0; environmental sustainability was slightly above 0 on average while social and economic sustainability were slightly below 0. In sections 4.1 to 4.3 we present the distributions of the IRR scores for each sustainability aspect, and in section 5 we offer insights about these results.

4.1. Social sustainability

The mean IRR scores and standard deviations for social sustainability are presented in Table 1.

[1] https://github.com/ndehaibi/krippendorff-alpha-irr

Table 1. Mean IRR scores and standard deviations for social sustainability

	Review Count	IRR Mean	IRR Standard Deviation
Letters	449	-0.114	0.848
Letters with NLP	441	-0.087	0.860
Words	449	-0.111	0.853
Words with NLP	441	-0.046	0.813

Histograms of social sustainability IRR scores for each Krippendorff's U-alpha measure are shown in Figure 3.

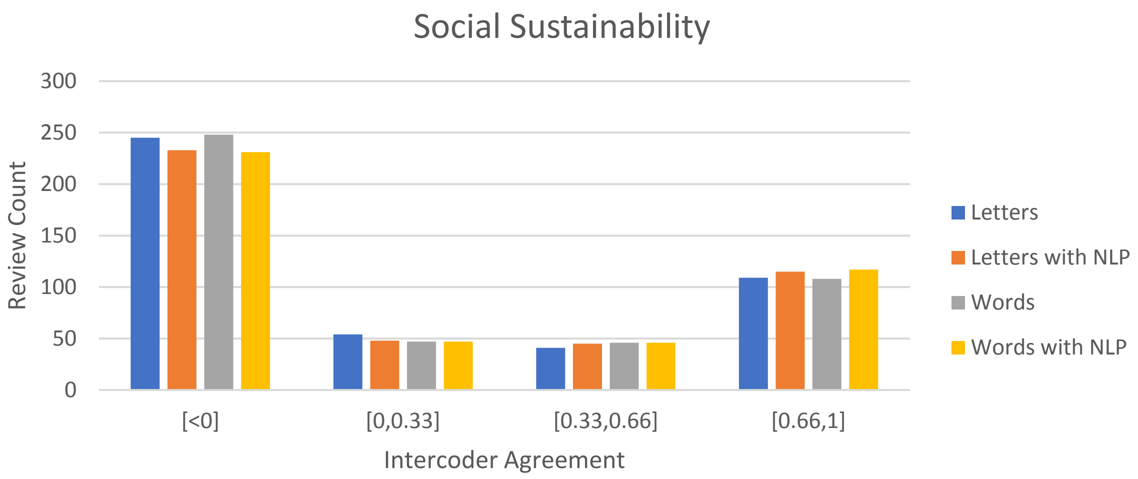

Figure 3. IRR for social sustainability

From Table 1, we can see that the review counts change from 449 to 441 when they are pre-processed with natural language processing. This is because some reviews may have annotations that contain only numbers or stop words; pre-processing in these cases would remove the annotation entirely. Table 1 also shows that the standard deviations are large relative to the mean. This is demonstrated by the distributions in the histograms shown in Figure 3 of IRR scores for each Krippendorff's U-alpha measure. Despite a mean score of around 0, the IRR scores for social sustainability range from about -3 to 1.

The histograms follow a similar distribution for all the Krippendorff's U-alpha measures; there is a spread from scores below 0 to 1 with the highest count of reviews being closer to 1. We see slight improvements in scores with measures that include NLP.

4.2. Environmental sustainability

The mean IRR scores and standard deviations for environmental sustainability are presented in Table 2.

Table 2. Mean IRR score and standard deviations for environmental sustainability

	Review Count	IRR Mean	IRR Standard Deviation
Letters	404	0.042	0.773
Letters with NLP	399	0.067	0.814
Words	404	0.0498	0.757
Words with NLP	399	0.097	0.779

Histograms of IRR scores for each Krippendorff's U-alpha measure are shown in Figure 4. While the scores are higher here than social sustainability, the distributions are very similar.

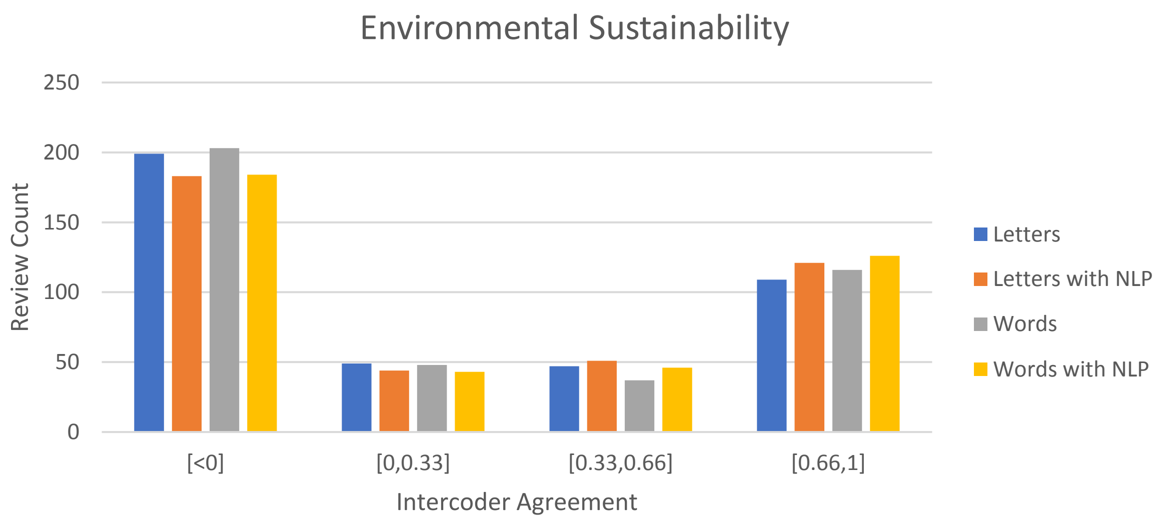

Figure 4. IRR for environmental sustainability

4.3. Economic sustainability

The mean IRR scores and standard deviations for environmental sustainability are presented in Table 3.

Table 3. Mean IRR score and standard deviations for economic sustainability

	Review Count	IRR Mean	IRR Standard Deviation
Letters	436	-0.088	0.905
Letters with NLP	433	-0.081	0.920
Words	436	-0.062	0.882
Words with NLP	432	-0.033	0.865

Histograms of IRR scores for each Krippendorff’s U-alpha measure are shown in Figure 5. Again, we see a very similar distribution compared to the other two sustainability aspects.

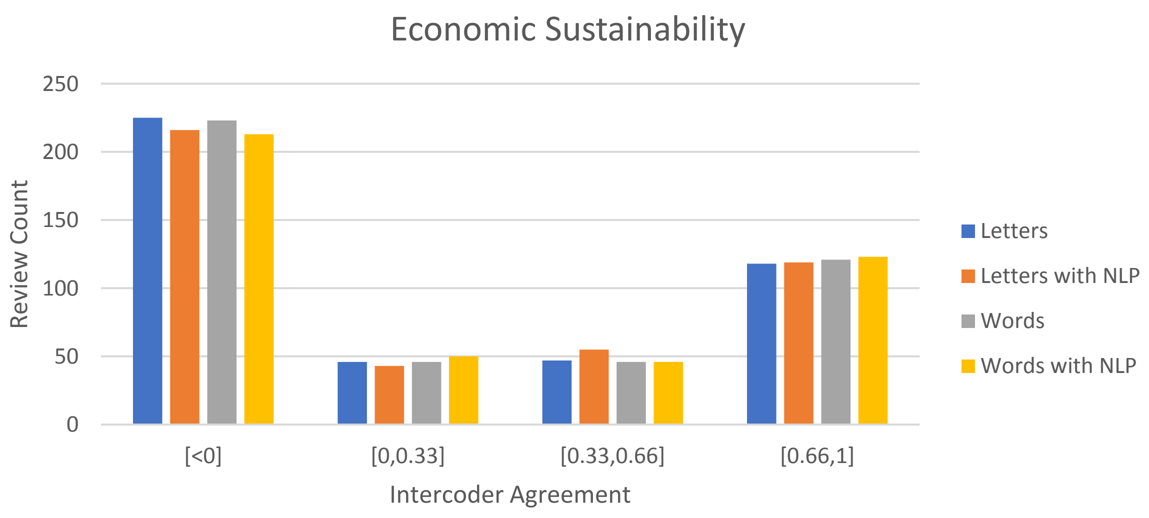

Figure 5. IRR for economic sustainability

5. Discussion

In light of these results, we present a discussion on how designers can use IRR scores in the context of assessing reliability of qualitative text annotations for machine learning datasets. The Krippendorff's U-alpha variations we implemented did not have a large effect on the IRR scores and so we examined the annotations that had the lowest IRR scores to better understand the results. Below is an example of one of these annotations that received an IRR score of -3:

- **Review:** Did not last a month of light use (every other day or so). The plastic nub that holds the strainer in place broke and now it's useless.
- **Annotator 1 highlight:** The plastic nub that holds the strainer in place broke and now it's useless.
- **Annotator 2 highlight:** Did not last a month of light use (every other day or so).

In this example, the first annotator highlighted the second half of the review while the second annotator highlighted the first half of the review. From the lens of Krippendorff's U-alpha, these annotations have no overlap and span different halves of the overall review. This suggests there are systematic differences between the annotators. Looking at this pair of annotations however we see that, semantically, both annotations revolve around durability of the product. The first sentence is a general statement about the durability, while the second provides more detail to explain the first statement. Therefore, there is some redundancy in the review and the annotations might be more closely related than IRR suggests. For this reason, we suggest that having a low Krippendorff's U-alpha score in the context of qualitative annotations for machine learning may not necessarily reflect a low agreement between the annotators.

We also propose that having a high agreement score may not be desirable or effective when building an annotated dataset for machine learning. Referring back to the annotation example above, we see that one annotator highlighted the first half of the review which was a general comment, while the other annotator highlighted the other half which was more specific. If both annotators had highlighted the general phrase, we would not have gained as much useful information despite having a higher agreement score. Therefore, in the context of machine learning, we suggest that the annotation task becomes more effective as a hunting exercise where we collect as much relevant information as we can. This is particularly the case with NLP tools like term-frequency inverse-document-frequency (TF IDF) that can reduce the importance of redundant terms and increase the importance of unique and specific terms in models. TF IDF is the product of term frequencies and inverse document frequencies (Equation 11) (Jurafsky and Martin, 2017).

$$w_{ij} = tf_{ij} * \log\left(\frac{N}{df_i}\right) \tag{11}$$

Equation (11) shows the TF IDF weight w_{ij} for word i in document j where N is the total number of documents and df_i is the number of documents where the word i occurs. The TF IDF transformation gives a higher weight to words that occur only in a few documents. Therefore, having a high agreement score becomes less important in this context when machine learning can mitigate annotations with less useful information while also benefiting from a larger dataset.

Particularly with qualitative topics like sustainability, it is expected that people will have different perspectives even if annotation training is provided. While a high IRR score may not be desirable in this context, we suggest that IRR can still provide useful information for designers. Looking at Figure 2, we see that on average environmental sustainability has a higher IRR than the other two sustainability aspects. This could suggest that annotators have a slightly more united perspective on what environmental sustainability is compared to social and economic sustainability. Environmental sustainability is generally the more prevalent aspect of sustainability and users may be more familiar with their perception of it, therefore reducing redundancies in reviews. This can inform designers on how they chose to design products involving environmental aspects. Moreover, looking at the histogram distributions in Figures 3 to 5, we see that there are four buckets of IRR scores ranging from below 0 to 1. These distributions could be useful to designers as they cluster annotations with high agreement and lower agreement, therefore helping designers identify perceptions that are more prevalent and perceptions that are more niche (perceptions on sustainability in this case). For example, using the clusters designers could identify sustainability perceptions that have a general consensus among customers, or focus on smaller market segments by looking at disagreements in the clusters.

Coming back to measuring reliability of qualitative annotations in machine learning datasets, we suggest that external validity metrics like accuracy, precision, and recall of the model are more effective measures despite being foreign in the design research space. To calculate these metrics, we would split the data into training, validation, and test sets and train the model to make sure that it is working and outputting results as expected (Jurafsky and Martin, 2017). Based on the metric scores of the model we would then be able to infer it the annotation dataset is reliable.

6. Conclusion

The goal of this study is to help designers measure reliability of qualitative annotations in the context of machine learning datasets using metrics that are common in the design research space. We investigated inter-rater reliability (IRR) as an internal validity measure by leveraging annotations of text data from a previous study where annotators highlighted social, environmental, and economic aspects of sustainability in online product reviews of French Press coffee makers. We calculated IRR scores of the annotations using four variations of Krippendorff's U-alpha: the first is the baseline where we looked at letter counts to measure differences between annotations, the second is where we looked at word counts to measure differences, and the third and fourth are the same as the first and second but with natural language processing of the reviews. The purpose of the variations was so that we may tune how sensitive the IRR measures are in different annotation scenarios.

We found that, while the variations slightly increased IRR scores from the baseline, the IRR scores on average ranged between -0.1 to 0.1. We examined annotations with the lowest scores and found that a low IRR score in the context of qualitative annotations may not necessarily reflect a low agreement between annotators due to potential redundancies in semantics. Moreover, we found many examples where, despite having low IRR scores, the annotators still captured useful information. In the case of machine learning datasets, we suggest that having a low IRR score might be preferable over high agreement between annotators to provide more unique data for a model to learn from. We discussed how this is particularly the case when tools like TF IDF can help balance for annotations with less useful content. Based on the results we propose that IRR can still be useful for designers in this context by clustering customer perceptions based on how well users agree or disagree on them. In terms of measuring reliability of this type of dataset, we propose that using external validation metrics like accuracy, precision, and recall are a better indicator of data quality despite them being foreign in design research. The results and discussions in this study are limited to the context of highly qualitative annotation tasks that are used as machine learning datasets.

References

Card, D. et al. (2015), "The Media Frames Corpus: Annotations of Frames Across Issues", *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*, Beijing, China, July 26-31, 2015, Association for Computational Linguistics, pp. 438-444. https://doi.org/10.3115/v1/P15-2072

Cohen, J. (1960), "A coefficient of agreement for nominal scales", *Educational and Psychological Measurement*, Vol. 20 No. 1, pp. 37-46. https://doi.org/10.1177/001316446002000104

El Dehaibi, N., Goodman, N.D. and MacDonald, E.F. (2019), "Extracting customer perceptions of product sustainability from online reviews", *Journal of Mechanical Design*, Vol. 141 No. 12, p. 121103. https://doi.org/10.1115/1.4044522

Fleiss, J.L. (1971), "Measuring nominal scale agreement among many raters", *Psychological Bulletin*, Vol. 76 No. 5, pp. 378-382. https://doi.org/10.1037/h0031619

Goodman, J.K. and Paolacci, G. (2017), "Crowdsourcing Consumer Research", *Journal of Consumer Research*, Vol. 44 No. 1, pp. 196-210. https://doi.org/10.1093/jcr/ucx047

Gwet, K.L. (2014), *Handbook of inter-rater reliability*, Advanced Analytics, Gaithersburg.

Hallgren, K.A. (2012), "Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial", *Tutor Quant Methods Psychol*, Vol. 8 No. 1, pp. 23-34.

Jurafsky, D. and Martin, J.H. (2017), "Naïve Bayes and sentiment classification", *Speech and language processing*, Stanford University.

Kennedy, L. et al. (2019), "Evaluation of a mindfulness-based stress management and nutrition education program for mothers", *Cogent Social Sciences*, Vol. 5 No. 1, pp. 1-12. https://doi.org/10.1080/23311886.2019.1682928

Krippendorff, K. (2004), "Measuring the reliability of qualitative text analysis data", *Quality and Quantity*, Vol. 38 No. 6, pp. 787-800. https://doi.org/10.1007/s11135-004-8107-7

Krippendorff, K. (2018), "Reliability", In: Accomazzo, T., Helton, E., Olson, A. and Ponce, M. (Eds.), *Content analysis*, Sage, Thousand Oaks, pp. 277-360.

Lai, V.K., Li, J.C. and Lee, A. (2019), "Psychometric validation of the Chinese patient- and family satisfaction in the intensive care unit questionnaires", *Journal of Critical Care*, Vol. 54 No. December 2019, pp. 58-64. https://doi.org/10.1016/j.jcrc.2019.07.009

Liang Y. et al. (2019), "Using social media to discover unwanted behaviours displayed by visitors to nature parks: comparisons of nationally and privately owned parks in the Greater Kruger National Park, South Africa", *Tourism Recreation Research*. https://doi.org/10.1080/02508281.2019.1681720

Paolacci, G. and Chandler, J. (2014), "Inside the Turk: Understanding Mechanical Turk as a Participant Pool", *Current Directions in Psychology Research*, Vol. 23 No. 3, pp. 184-188. https://doi.org/10.1177/0963721414531598

Rash, J.A. et al. (2019), "Assessing the efficacy of a manual-based intervention for improving the detection of facial pain expression", *European Journal of Pain*, Vol. 23 No. 5, pp. 1006-1019. https://doi.org/10.1002/ejp.1369

Stab, C. and Gurevych, I. (2014), "Identifying argumentative discourse structures in persuasive essays", *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, October 25-29, 2019, Association for Computational Linguistics, pp. 46-56. https://doi.org/10.3115/v1/D14-1006

Stone, T. and Choi, S.K. (2013), "Extracting consumer preference from user-generated content sources using classification", *Proceedings of the ASME 2013 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Portland, OR, August 4-7, 2013, Association of Mechanical Engineers, pp. 1-9. https://doi.org/10.1115/DETC2013-13228

Toh, C.A., Miller, S.R. and Kremer, G.E. (2014), "The Impact of Team-Based Product Dissection on Design Novelty", *Journal of Mechanical Design*, Vol. 136 No 4, p. 041004. https://doi.org/10.1115/1.4026151

Tuarob, S. and Tucker, C.S. (2015), "Automated discovery of lead users and latent product features by mining large scale social media networks", *Journal of Mechanical Design*, Vol. 137 No. 7, p. 071402. https://doi.org/10.1115/1.4030049