Insights into developing analytical categorization schemes: three problem types related to annotation agreement

Pihla Toivanen
University of Helsinki

Eetu Mäkelä
University of Helsinki

Antti Kanner
University of Turku

Abstract

Coding themes, frames, opinions and other attributes are widely used in the social sciences and doing that is also a base for building supervised text classifiers. Coding content needs a lot of resources, and lately this process has been utilized particularly in the training set annotation for machine learning models. Although the objectivity of coding is not always the purpose of coding, it helps in building the machine learning model, if the codings are uniformly done. Usually machine learning models are built by first defining annotation scheme, which contains definitions of categories and instructions for coding. It is known that multiple aspects affect the annotation results, such as, the domain of annotation, number of annotators, and number of categories in annotation. In this article, we present few more problems that we show to be related with the annotation results in our case study. Those are negated presence of a category, low proportional presence of relevant content and implicit presence of a category. These problems should be resolved in all schemes on the level of scheme definition. To extract our problem categories, we focus on a media research case of extensive data on both the process as well as the results.

1 Introduction

The coding of content features such as themes, frames, opinions and so on are widely used in the social sciences. In essence, the purpose of coding is to turn unstructured (qualitative) data such as text into structured data (the codes and their appearances) on which inferences can be made (King et al., 2021). The purpose of this study is to present few more characteristics, that has been already known, of the texts that cause difficulties in human-made coding, in other words, annotation.

Doing coding by hand is a resource-intensive process, particularly at scale (Beresford et al., 2022). Thus, within the computational social sciences, there have been many efforts to enable algorithms to do the coding for us (Macanovic, 2022; Grimmer et al., 2021).

Here, two distinct approaches appear, targeting different styles or phases of coding. The first is to use an unsupervised approach, such as clustering or topic modelling for an initial coding of the data, used to get an initial at-scale understanding of it (Isoaho et al., 2021; Grimmer et al., 2021; Macanovic, 2022). The second aligns with the more consolidated, theory- or hypothesis-informed codes which form a basis for inference. Traditionally, these codes are often created in a second pass of coding based on information gained from the initial codes, or alternatively may already be defined at the beginning of research that employs a ready theory or hypothesis. Here, the dominant mode of operation in computational social sciences currently is to train a classifier based on manual training data.

While not completely obviating the need to do manual coding, the core idea here is that only a small portion of the data overall needs to be coded, and the classifier will handle propagating the codes to the rest of the material. This in turn is argued to lead to the possibility of using much larger unstructured datasets as research material, which is also argued to both broaden as well as solidify the inferences that can be made based on them.

Producing training data annotations for building text classifiers has been studied previously from various perspectives of the coding or annotation process: instructions given to the annotators (Budak et al., 2021), the number of tasks given at the same time (Finnerty et al., 2013), and the text difficulty of the classified texts (Weber et al., 2018). Finnerty et al (Finnerty et al., 2013) found that simplicity and the amount of tasks affects the agreement between annotators. Weber et al (Weber et al., 2018) found that text difficulty, measured by lexical diversity measure type-token ratio, predicts intercoder reliability so that increasing the diffi-

Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities, pages 570–577

culty lowers the reliability. Budak et al (Budak et al., 2021) found that training of the annotators improves the annotation results while using codebook not. Also various other features have been shown to affect to the annotation performance, for example, annotation domain, number of annotators and number of annotation categories (Bayerl and Paul, 2011).

Detecting annotation errors in general, not only for full texts, has been studied for various tasks, such as slot filling (Larson et al., 2020), and part-of-speech tagging (Kveton and Oliva, 2002). Annotation error detection can be divided to two areas: detecting them based on statistical measures and based on grammatical comparison (Dickinson, 2015). Statistical error detection relies on finding anomalies from the annotations, such as rare local tag patterns in linguistic annotations (Eskin, 2000). Examples of grammatical comparison are pattern matching to detect invalid bigrams from a POS tag sequence (Kveton and Oliva, 2002), and utilizing different layers of linguistic information to find inconsistencies between the layers, such as POS tags and syntactic tree, and using them to correct erroneous POS tags (Hockenmaier, 2003).

In the annotation error detection approaches above, ground truth has been known. For example in part-of-speech tagging, it is possible to define the correct tags, where in more complex tasks, identifying annotation errors is difficult as the correct class cannot be determined. There are also ways to take into account disagreement between the annotators in building the classification model, which is useful especially when the ground truth labels are not known. It can be done by training the classifiers using annotation distributions (Peterson et al., 2019), adjusting the label distribution based on removing noise (Gordon et al., 2021), or with jury learning, where annotators are chosen to compose a jury, that can be formed based on external attributes, such as to have balance between different genders (Gordon et al., 2022).

In current practice, the application and validation of machine-learned classifiers for coding usually happen in distinct stages (Krippendorff, 2019). First, to evaluate both the coding scheme as well as annotation quality, typically a measure of inter-annotator agreement such as Cohen's Kappa (Cohen, 1960) or in the case of multi-class and possibly multi-label settings, Krippendorff's Alpha (Krippendorff, 2011) is used. If this is good enough, the process moves to training the classifier and evaluating it. Most often this is done using the standard computer science evaluation metrics of precision and recall, as well as their harmonic mean, the F1-score. In a multi-class setting, performance measures across classes are also often summarised into a single number using either micro- or macro-F1 measures, although both have drawbacks (Harbecke et al., 2022). Finally, once the classifier accuracy is deemed good enough, the most common approach is to just use the numbers produced by the classifier as is, although ways have also been designed to factor in classifier biases (Hopkins and King, 2010; Bachl and Scharkow, 2017).

A glaring problem in the workflow described above is that the evaluation of all of coding scheme, annotation quality and classifier accuracy happen in isolation from each other, and most often, none of the uncertainty identified at each stage is carried forward (Grimmer et al., 2015; Song et al.; Bachl and Scharkow, 2017). In the field of qualitative methods, it has been even questioned, how realistic or desirable end result complete objectivity of annotation is (O'Connor and Joffe, 2020), and that is a problem for training machine learning classifiers.

In practice, this can lead to the final data used for inference widely diverging from what its users expect. However, in this paper, we will not be discussing how to explicitly link these stages. Instead, we will focus on trying to lessen the uncertainty in the first place. We start from the first stage of the process, the definition of the coding categories, and add three more features in the texts, in addition to those that have been already known, that lead towards increasing fuzziness and uncertainty, worse inter-annotator agreement and analytical usefulness.

To extract our three problem categories, we focus on a research case of extensive data on both the process as well as the results.

2 Case description

The study is based on experiences and annotation scores from an annotation project that sought to categorize Finnish news media texts concerning alcohol policy. The aim of the annotation project is to develop a training dataset for a supervised classifier that detects categories related to Finnish alcohol policy discussion to be used in a study of Finnish political journalism. While the results of the aforementioned study will be published in due

course, this article discusses issues related to the annotation process and, more specifically, provides a detailed analysis on how the annotation disagreement is distributed across articles of the data. The main goal is to give insight on how the annotation scheme performed in the context of annotation process. The data for this article are the inter-annotator scores of each annotated article.

The annotation scheme is based on an earlier Finnish study on representations of alcohol policy in Finnish news media. This scheme was iteratively and reactively developed further to enhance the initially unsatisfactory inter-annotator scores. The iterative developments in the categorisation scheme included modifications both in the category level, such as dropping some categories from the original scheme, and excluding articles related to foreign issues. All changes to the annotation scheme were done in order to bring the annotation closer to the media studies aims of the study. The main motivation for the changes was to make the classification suit better the research interests and to improve the inter-annotator scores by dropping categories that were intuitively deemed as tricky for annotators. Thus, the annotators were finally instructed to categorize the articles in three categories:

1. Alcohol legislation: regulation of Finnish alcohol markets and availability of alcohol beverages, reforming Finnish alcohol act. Articles about local crimes or occurrences of crime committed under the influence of alcohol were excluded from the definition.

2. Alcohol markets and alcohol consumption research: Statistical reports about alcohol sales, alcohol consumption or people drinking alcohol in different situations

3. Alcohol harms and their prevention and treatment: Social, health and public disorder problems caused by alcohol consumption, a service system that reduces alcohol problems, multi-professional activities aimed at preventing alcohol problems, such as education, organization work, youth work and police surveillance

3 Dataset and inter-annotator tests

The base dataset of the annotation project included in total 33,902 articles from four Finnish news media: Helsingin Sanomat, Yle, Iltalehti and STT.

Table 1: Annotation rounds

	Round		
	1	2	3
N (articles)	50	50	100
Annotation scheme	single	multi	multi
Inter-annotator score	0.68	0.75	0.64

The base dataset included all articles mentioning the Finnish lexeme for alcohol, “alkoholi” and all of its word forms. The analysis presented in this article are based on three annotation tests (Table 1), the articles of which have been randomly selected from the base dataset. In addition to creating annotation tests, we also formed a dataset of 1,500 articles, each of which had three annotations, for the aim of text classifier training. The last annotation test (Annotation round 3) was part of that effort.

On the first annotation round (Table 1), the categories were defined in a single label scheme where the annotator should choose one of the categories or “not alcohol policy”. Annotators were also divided into two groups, where one of the groups had a fifth category, “alcohol as a side note”, as an annotation option. On the second and third annotation rounds annotation was done by 10 students in a multi-label scheme with options “not alcohol policy” or 1-3 of the categories described above.

4 Ruling out annotator effects

After noticing lower inter-annotator scores in the annotation project than in the preceding annotation tests, possible annotator effects were studied in detail. We found one outlier annotator (annotator 4 in Figure 1), whose performance in the pairwise agreement plot was visibly different than the performance of other annotators. The underlying reason for this was one annotator who failed to annotate STT’s news, although submitted all annotations in the annotation user interface as empty annotations.

After the removal of STT annotations of the annotator 4, there was still a disagreement in the annotations, which can be seen in Figure 2. The remaining potential sources of error were therefore the classification framework and the article texts.

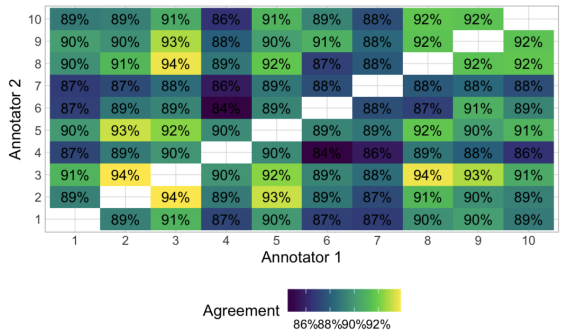

Figure 1: Pairwise annotator agreements

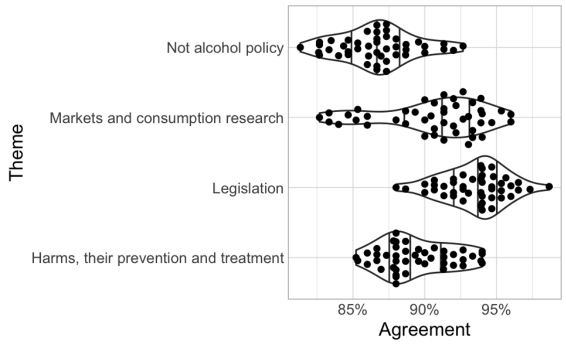

Figure 2: Label-wise inter-annotator agreements

5 Surfacing effects of how phenomena appear in the data

5.1 Extracting problem categories through close reading

After excluding clear outlier annotations, we were interested in investigating how choices made in developing our annotation scheme and the features of the articles contributed to disagreements among annotators. Through iterative close reading of both the articles and annotations of each three annotation rounds, we identified three properties of articles in relation to categories in the annotation scheme that tentatively seemed to be connected to high disagreement scores. All three properties relate to how the category-relevant content is positioned in the articles. We will discuss the properties in what follows.

5.2 Low proportional presence of relevant content

In the annotation scheme document of the project, it was explicated that if content related to a given category was present in the text at all, it should be annotated in that category. Our results show that in practice the annotators follow this instruction to only a degree. There seems to be a threshold for the presence of category-relevant content in the article below which the annotators are more likely to disagree on the category. This threshold can concern either the coverage of the relevant span or its focality. That is to say, it is about either how much of the word count of the text is directly category-relevant, or how central, focal or important the category-relevant content is from the point of view of the article as a whole. If either of these is low, the category can be considered a low proportional presence in the article. Often these two modes are mutually dependent. Consider a typical example of an article from Yle about Finnish health care officials experiences in dealing with Russian insurance companies' policies in cases ofRussian tourists requiring health care in Finland. Within the relatively long article, alcohol is mentioned as an anecdote of Russian insurance companies' compensation terms, where compensation is rarely given if alcohol has been involved in the incident (18.1.2011, titled "Collection of Russian treatment costs requires expertise"). Here, the amount of relevant content is proportionally small and extraneous to the topic of the article.The annotators disagreed in the categorization of the article . However, there are examples where the category-relevant content has only limited span of coverage but has a very focal role in the text. These are common for example in cases where the article contains multiple quotations from different people (politicians, experts, officials and so on) and the category-relevant span is in one of the quotations. In cases like this, the perspective provided by one quotation can be very important in media studies perspective, for example when a doctor provides a medical perspective or legal expert a constitutional one. A major practical issue regarding these two modes of low proportional presence, is that while the proportional coverage can be measured by comparing the length of the relevant span to the text as a whole, the latter is quite subjective and contextual.

5.3 Implicit presence

Many of the articles with high disagreement had an irregularity on how the relevant content was present in them. In many of them, the concepts related to the annotated category were named or referred to implicitly rather than explicitly. These implications could be based on conventional conceptual relations such as conceptual hierarchies (hyponym

hyperonym), part-whole relations, cause-effect relations or other associative relations the annotators recognized. The agreement between annotators thus required making the same implicit connections between things. For instance the annotators disagreed on whether a mention of police as an implied reference to alcohol legislation or serving alcohol to minors was a marker of the presence of alcohol harms. This highlights the fact that often times, like in our case, the categories themselves can be related. An article discussing serving alcohol to minors could easily be categorized under both alcohol harms (because alcohol is explicitly harmful to minors in clinical sense) and alcohol legislation (because it is illegal). A similar relation can be identified in the case of driving under the influence of alcohol: it is widely accepted as a liability to traffic safety and would thus fall under the category of harms and because it is criminal, it has clear legal implications. These relations between categories are brought about by the fact that alcohol legislation is often based on conceptions of harms of alcohol. Thus, there is a cause-effect relation between alcohol harms and alcohol legislation and, consequently, reference to one easily points to the other as well. A concrete example of implicit presence is an article about a study on how hot drinks may cause cancer (Iltalehti 15.6.2016, titled "Very hot drinks can cause cancer"). Most of the article is about reporting results of recent research which observed that high temperature of a drink may cause esophageal cancer, but there is also a mention of alcohol consumption being controlled in the research setting. Most of the annotators annotated the article to the category "alcohol harms" even though the article did not take a position on whether alcohol reduces or increases the risk of cancer. The connection between health risks and alcohol was based on annotators' encyclopedic knowledge and their familiarity with conventions of this type of journalism. Based on our qualitative analysis, the occurrence of implicit cues as a basis of category annotation is a comparatively rare phenomenon. By analyzing 150 articles, the dataset with 100 annotations from the final project and 50 articles from the second annotation test before the final project, we found in total 20 articles with implicit cues of one or more of the categories in the text. We observed that the proportion of same annotations among the annotators was lower in the case of articles with implicit cues of the categories, indicating that with implicit presence of the category, it may be more difficult for the annotators to agree on categorization.

5.4 Negated presence of category

Another common feature for articles with high disagreement was the presence of category-relevant content in a negated form. Generally, negation fell under two distinct categories. In the first category there were articles, where the relevant concepts were referred to by their opposites, either lexical opposites or more complicated diametrically oppositional propositional structures. For example, a news article from STT (5.1.2006, titled "Every fifth Finnish do Dry January") discusses/reports people spending the whole of January abstaining from alcohol, citing health benefits as their main motivation. The article is relatively short, only 5 sentences. The alcohol harms, then, are not referred to directly, but by through the antonymical relation between health and harm. Thus, the annotators disagreed over whether the proposition that "abstaining from alcohol has health benefits" constitutes a reference to harms of alcohol consumption or not. In the second type, the relevance of a given perspective (to which a category membership is linked in the annotation scheme) is explicitly denied. Similar cases are the ones where the text explicitly takes into account the reader's expectations and contradicts them. A common convention in Finnish news reporting especially in cases of traffic accidents or crimes of violence is to note whether or not alcohol was involved in the incident. When the involvement of alcohol is denied, the reader's expectations concerning probable involvement are simultaneously acknowledged and enforced by implying that this is a type of situation where alcohol usually is involved. The annotation result of the article in question included one "not alcohol policy" annotation, five annotations in alcohol harm category and eight annotations in "alcohol markets and alcohol consumption research". Because the article was only five sentences long and the only topic was spending January without drinking alcohol, we interpret that the reason for half of the annotators agreeing and the other half not agreeing in the alcohol harms category was the opposite presentation of the category in the text. There were two other articles about voluntary sobriety and one article about the unrelatedness of alcohol in an accident with a similar annotation profile than the example

article above.

6 Effects of the problem types

After tentatively identifying the three factors above contributing to disagreement , we sought to measure how much they could explain the disagreement between annotators. To answer this question, we annotated altogether 150 articles, consisting of 100 from the final project and 50 from the second annotation test, using the three problem categories above. Annotation was done using a binary classification and each article was annotated whether it had a proportional presence of some category, whether it had an implicit presence of a category and whether it had a negated presence of a category. In addition, we calculated the overall properties of the articles and student annotations for each article:

1. Percent of articles with complete annotation agreement among the article group
2. Mean of majority theme proportions among the article group: we calculated a majority theme from the annotations and proportion of that for each article, and then calculated the mean of that value for both article groups.
3. Mean of annotation time centiles among the article group: because of outlier annotations with unrealistically large annotation time values (probably the annotator had taken a break and left the annotation user interface open), we present all annotation times as 10-centile values. In that way we can safely calculate the mean of the values without letting the large values to have too much effect on the result.
4. Mean of text lengths among the article group

We compared the articles where the problem categories were present to the articles where the categories were not present in four characteristics described above.

In general, those articles where one of the above mentioned features were present had much lower level of agreement that the ones with none (Tables 2, 3, and 4). For all three, the annotation time was longer in articles where they were present. In the case of a low proportional presence, longer annotation time compared to normal or higher proportional presence was naturally explainable by their longer article length (Table 2). When the articles had implicit or negated presence category markers, the length of the article was actually shorter, while the annotation time was still longer (Table 3, Table 4). This indicates that the reason for the longer annotation time could be related to difficulties in applying the annotation scheme in cases where the markers are present in the article in an atypical form.

7 Discussion

In this article, we have identified few problem types that we believe pervade many annotation schemes: low proportional presence of a category, implicit presence of a category and negated presence of a category. We have shown that in our case study, presence of these problem types in articles lead to lower levels of agreement. Based on our empirical results, we have a few recommendations for projects that involve creating annotation schemes. First, it is important to make explicit choices with regard to wanting to extract "a clearly present category" vs "a category that appears in any measure" (majority vs minority theme e.g. still problematic, because leaves option for "appears in a side sentence" to be declared majority theme if nothing else appears). Second, no annotation principle or category definition should depend on the presence or absence of any other category. Designing categories should be done so that it should be possible to annotate each category in isolation. Third, after a round of inter-annotator-agreement, the focus should be on disagreements, but not in isolation but as a whole. The following question should be asked: What unites the articles with disagreements? How does this contrast with commonalities in the articles/categories without disagreements? We argue that with focusing on disagreements before finishing the annotation scheme, the quality of the final scheme would be better.

Limitations

We agree that this study concerns only one case, alcohol policy media articles. However, we believe that the problem categories identified are useful in multiple cases and can be found in other text types too.

Acknowledgements

This work was supported by the Academy of Finland under Grant number 320677.

Table 2: Agreement statistics of articles with gradational presence of a category

	Gradational presence	Only non-gradational presence
N =	37	113
N (annotations) =	359	1106
Proportion of articles with complete agreement	0	0.6
Mean majority theme proportion	0.67	0.89
Mean annotation time centile	6.66	5.11
Mean content length	2674.62	1736.27

Table 3: Agreement statistics of articles with implicit presence of a category and only explicit presence of categories

	Implicit presence	Only explicit presence
N =	23	127
N (annotations) =	226	1239
Proportion of articles with complete agreement	0.04	0.53
Mean majority theme proportion	0.68	0.87
Mean annotation time centile	5.74	5.44
Mean content length	1686.304	2018.701

Table 4: Agreement statistics of articles with opposite presence of a category and only direct presence of categories

	Opposite presence	Only direct presence
N =	6	144
N (annotations) =	59	1406
Proportion of articles with complete agreement	0	0.47
Mean majority theme proportion	0.75	0.84
Mean annotation time centile	5.97	5.47
Mean content length	1562.3	1984.63

References

M. Bachl and M. Scharkow. 2017. Correcting measurement error in content analysis. *Communication Methods and Measures*, 11(2):87–104.

P.S. Bayerl and K.I. Paul. 2011. What determines inter-coder agreement in manual annotations? a meta-analytic investigation. *Computational Linguistics*, 37(4):699–725.

M. Beresford, A. Wutich, M. V. du Bray, A. Ruth, R. Stotts, C. SturtzSreetharan, and A. Brewis. 2022. Coding qualitative data at scale: Guidance for large coder teams based on 18 studies. *International Journal of Qualitative Methods*, 21:16094069221075860.

C. Budak, R. K. Garrett, and D. Sude. 2021. Better crowdcoding: Strategies for promoting accuracy in crowdsourced content analysis. *Communication Methods and Measures*, 15(2):141–155.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

M. Dickinson. 2015. Detection of annotation errors in corpora. *Language and Linguistics Compass*, 9(3):119–138.

E. Eskin. 2000. Automatic corpus correction with anomaly detection. In *Proceedings of the First Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-00)*, pages 148–153, Seattle, Washington.

A. Finnerty, P. Kucherbaev, S. Tranquillini, and G. Convertino. 2013. Keep it simple: Reward and task design in crowdsourcing. In *ACM International Conference Proceeding Series*, pages 2–5.

M. L. Gordon, M. S. Lam, J. S. Park, K. Patel, J. Hancock, T. Hashimoto, and M. S. Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *CHI Conference on Human Factors in Computing Systems (CHI '22)*, pages 1–19.

M. L. Gordon, K. Zhou, K. Patel, T. Hashimoto, and M. S. Bernstein. 2021. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*, pages 1–14.

J. Grimmer, G. King, and C. Superti. 2015. The unreliability of measures of intercoder reliability, and what to do about it.

J. Grimmer, M. E. Roberts, and B. M. Stewart. 2021. Machine learning for social science: An agnostic approach. *Annual Review of Political Science*, 24(1):395–419.

D. Harbecke, Y. Chen, L. Hennig, and C. Alt. 2022. Why only micro-f1? class weighting of measures for relation classification. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 32–41, Dublin, Ireland.

J. Hockenmaier. 2003. *Data and models for statistical parsing with Combinatory Categorial Grammar*. Ph.D. thesis, School of Informatics, The University of Edinburgh.

D. J. Hopkins and G. King. 2010. A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1):229–247.

K. Isoaho, D. Gritsenko, and E. Mäkelä. 2021. Topic modeling and text analysis for qualitative policy research. *Policy Studies Journal*, 49(1):300–324.

G. King, R. O. Keohane, and S. Verba. 2021. *Designing social inquiry: Scientific inference in qualitative research*, new edition edition. Princeton University Press.

K. Krippendorff. 2011. Computing krippendorff's alpha-reliability.

K. Krippendorff. 2019. *Reliability. In Content analysis: An introduction to its methodology (pp. 387–420).* SAGE Publications, Inc.

P. Kveton and K. Oliva. 2002. (semi-)automatic detection of errors in pos-tagged corpora. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 107–113.

S. Larson, A. Cheung, A. Mahendran, K. Leach, and J. K. Kummerfeld. 2020. Inconsistencies in crowdsourced slot-filling annotations: A typology and identification methods. In *Proceedings of the 28th International Conference on Computational Linguistics*.

A. Macanovic. 2022. Text mining for social science – the state and the future of computational text analysis in sociology. *Social Science Research*, 108:102784.

C. O'Connor and H. Joffe. 2020. Intercoder reliability in qualitative research: Debates and practical guidelines. *International Journal of Qualitative Methods*, 19:1609406919899220.

J. C. Peterson, R. M. Battleday, T. L. Griffiths, and O. Russakovsky. 2019. Human uncertainty makes classification more robust. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9617–9626.

H. Song, P. Tolochko, J.-M. Eberl, O. Eisele, E. Greussing, T. Heidenreich, F. Lind, S. Galyga, and H. G. Boomgaarden. In validations we trust? the impact of imperfect human annotations as a gold standard on the quality of validation of automated content analysis. *Political Analysis*.

R. Weber, J. M. Mangus, R. Huskey, F. R. Hopp, O. Amir, R. Swanson, A. Gordon, P. Khooshabeh, L. Hahn, and R. Tamborini. 2018. Extracting latent moral information from text narratives: Relevance, challenges, and solutions. *Communication Methods and Measures*, 12(2–3):119–139.