Whose Truth is it Anyway? An Experiment on Annotation Bias in Times of Factual Opinion Polarization *

Mariken A.C.G. van der Velden
Dep. of Communciation Science, Vrije Universiteit Amsterdam
Felicia Loecherbach
Amsterdam School of Communication Research, University of Amsterdam
Kasper Welbers
Dep. of Communciation Science, Vrije Universiteit Amsterdam
Myrthe Reuver
Computational Linguistics & Text Mining Lab, Vrije Universiteit Amsterdam
Antse Fokkens
Computational Linguistics & Text Mining Lab, Vrije Universiteit Amsterdam
Wouter van Atteveldt
Dep. of Communciation Science, Vrije Universiteit Amsterdam

Information shapes citizens' political decision-making. This process is amply studied by social scientists, who have human annotation as a crucial instrument in their toolkit. To address concerns regarding the validity and reliability of annotation tasks, we establish strict standards. Due to the democratization of data and the advances in NLP more data can be analyzed or classified, making these standards are more important than ever: An algorithm trained on biased data will reproduce and often exacerbate bias. Our tools to create valid and reliable annotation data currently do not allow for dealing with different perspectives of annotators. In two pre-registered experiments in the United States and the Netherlands, we show that personal characteristics of annotators, like political ideology or knowledge, interfere with annotators' judgement of political stances. Our results show that to improve annotated data for automated text analyses, and for stance detection models in particular, we need to critically evaluate how we create our gold standards.

Keywords: Experiment, Annotation Bias, Ideology, Measurinig Political Position, Text-as-Data, Political Knowledge

---

*Corresponding author: MACGvdV, Replication files are available on the author's Github account (https://github.com/MarikenvdVelden/bias-experiment); Author contributions: a) designed the study: MACGvdV, MR, AF, FL, WvA, & KW; b) conducted the study: MACGvdV, FL & MR; c) data cleaning & analysis: MACGvdV; d) writing of the paper: MACGvdV, MR, AF, FL, & KW

Introduction

Information shapes citizens' political decision-making. This process is amply studied by social scientists: Classical theories of political communication, such as agenda setting or framing (e.g. Van Aelst and Walgrave 2016; Lecheler and De Vreese 2019), formulate that political information drives opinion formation and participation in politics – from voting to protests. To gather and analyze the information citizens receive, the democratization of data and advent of computational social science has paved the way for new possibilities (for recent overviews, see Van Atteveldt, Trilling, and Calderon 2022; Grimmer, Roberts, and Stewart 2022). In particular, advances in Natural Language Processing (NLP) have made it possible to automatically analyze large quantities of data using machine learning (e.g., see Akyürek et al. 2020; Piskorski et al. 2023). To determine the validity of such large scale computational analyses, we rely on "gold standard" data, created by human annotators. It follows that the validity of these analyses hinges on the quality of these gold standards. Most data collection efforts to create gold standards assume that there is only one correct interpretation for every input example. Disagreement between the annotators is considered a bias – i.e. a systematic deviation between the "true" value of a theoretical concept in a population and how it is observed by the coder – and thereby something that needs to be dealt with at all costs (Aroyo and Welty 2015a). If a gold standard contains biases, it can foster bias in any downstream analysis.

Crowd-coding platforms have been shown to be reliable platforms to collect a gold standard (Van Atteveldt, Van der Velden, and Boukes 2021). Yet, from experimental studies (Clifford, Jewell, and Waggoner 2015; Huff and Tingley 2015; Berinsky, Margolis, and Sances 2014) we know that these online platforms are populated by people that are "unlike" the general public, being younger and holding more liberal values. This could mean that we introduce *sampling bias* into the creation of the gold standard if we stricly consider only one position to be true. Ennser-Jedenastik and Meyer (2018) report that coders of political texts

incorporate prior beliefs about parties' issue stances into their coding decisions. The authors find that party labels cue coders to a stance. For example, coders are more likely to report a left-wing party to be pro-immigration and a populist right-wing party to be against based on the exact same sentence. *Is this actually bias or a diversity of viewpoints? And how big of a problem is this bias/diversity of viewpoints?*

In this context, bias would occur if there is one objectively correct solution to a coding task and coders systematically get the answer *wrong* due to prior beliefs. Opposed to this, diversity of viewspoints rather indicates that there are actually multiple ways of assessing the situation which all can be (in some way, depending on the siutation) correct. There is a long history of text annotation in studies analyzing political text. Yet, the main underlying assumption that there is one correct interpretation for every input example remains untouched. However, especially for more complex coding tasks, research fields differ in their interpretations of what counts as a *correct* solution depending on the main goal of annotation. Do we want to create a clean benchmark data set that follows strict linguistic rules? Or do we want one that reflects how statements are being interpreted by "real" readers, how the perception of different types of respondents differs for the same stimulus? This could potentially affect downstream tasks as manual stance annotations are further used for e.g. sentiment analysis or frame analysis. Are we modelling the sentiment and frames of a particular section of the population while the results would change with a different sample? This question also taps into newer developments related to data annotations. Increasingly, Large Language Models such as ChatGPT are being used by researchers (Gilardi, Alizadeh, and Kubli 2023) as well as annotators themselves (Veselovsky, Ribeiro, and West 2023) to "simulate" human respondents for annotation tasks. These models draw among other things from the gold standard data sets that have been created by researchers, perpetuating the specific biases that are present in them and not allowing for any further disagreement or diversity of viewpoints.

Trends of polarization have shown that people do interpret information according to their

ideological position (e.g., Rekker and Harteveld 2022; Lee et al. 2021). Does this mean that some are right and others are wrong? Or is there an ideological difference in the ground truth? These questions present a fundamental challenge to the main way of working when collecting gold standard data, as we operate from the baseline assumption that disagreement among the annotators should be avoided or reduced. Typically, when specific cases continuously cause disagreement, more instructions are added to limit interpretations (for recent innovations to improve annotation, see for example: Barberá et al. (2021); Struthers, Hare, and Bakker (2020); Winter, Hughes, and Sanders (2020); DeBell (2013); Benoit et al. (2016); Ying, Montgomery, and Stewart (2022)). However, work in computational linguistics has shown that increased annotation instructions do not increase quality (Parmar et al. 2023). This leaves us between a rock and a hard place. *Is there a potential bias in annotators that we should account for?*

In this paper, we build upon the NLP literature on disagreement – or bias – in annotation (e.g., see Q. Shen and Rose 2021; Geva, Goldberg, and Berant 2019; Sommerauer 2020; Plank, Hovy, and Søgaard 2014) and so-called perspectivism (Cabitza, Campagner, and Basile 2023; Havens et al. 2022) – i.e. the adoption of methods that integrate the opinions and perspectives of the human subjects involved in the knowledge representation step of the machine learning processes (Cabitza, Campagner, and Basile 2023). We test the extent to which disagreement takes place and for what type of stances. In addition, we provide guidance on how to deal with diversity in conceptions and political heterogeneity when creating gold standard data. We use two high-powered pre-registered experiments (see here and here) in the Netherlands – a low-level polarized country – and in the U.S. – a high-level polarized country. We test possible biases stemming from ideology and political knowledge for correctly and overinterpreting stances. Moreover, based on the results of Ennser-Jedenastik and Meyer (2018), we test whether masking the political actor to mitigates these biases.[1] Another offered solution by Webb-Williams et al. (2023) is to collect information on po-

[1]The data and research compendium is published on the [main author's github page] – ommitted for the review process.

litical positions of the annotators to potentially re-weight responses based on population proportions to create composite measures, which is the focus of our H1 and H2. In the experiments, we vary the level of specification with which a political actor takes a stance – a declarative sentence versus a sentence where with some knowledge on politics, the stance might be inferred – as well as whether the political actor is shown or masked with putting `[ACTOR]` instead of the political party. We do this for four different political issues: *Environment*, *Immigration*, *Tax Policy*, and *EU* (for the Dutch case) or *Foreign Policy* (for the American case). This country selection not only allow us to showcase the scope conditions of disagreement due to different levels of political heterogeneity, but also differentiates between languages. English is not only the most dominant language for computational text analysis in social science (Baden et al. 2022; Dolinsky et al. 2023), crowd-coders do not need to be first language speakers, given the dominance of English in our daily lives. This is different for Dutch, it is a language spoken by a smaller community, typically first language speakers, yet still an often-enough researched case in computational text analysis in the social science (Baden et al. 2022; Dolinsky et al. 2023).

Overall, our results demonstrate that sentences where with some knowledge on politics the stance might be inferred are more likely to be overinterpreted, inferring a position where none is explicitly given. This is problematic as these sentences are very common in political text – like legislative debates or speeches – as well as media reports. Moreover, our results also demonstrate that for these disagreements in the crowd to occur, the context is important, as we find differences in disagreements between the American and Dutch context. Our findings underline the importance of taking disagreement seriously for the creation of gold standard text – the bread-and-butter of all machine learning endeavors. We should look beyond the majority vote and model it in the data, because if an algorithm is trained on biased data from disagreeing annotators, it will reproduce and often exacerbate that bias when it is applied to new data (e.g., see Prost, Thain, and Bolukbasi 2019). We thus rather opt for modelling a diversity of viewpoints instead of choosing for (potentially) biased solving of disagreements.

To be able to model these annotators' characteristics, we should survey the characteristics of annotators when using the crowd (see Webb-Williams et al. 2023 for a similar argument, yet different annotator characteristics).

Whose Truth is it Anyway? Disagreement & Perspectivism in Creating Gold Standard Data

Generating large data sets has become one of the main drivers of progress in NLP. Studying political texts, the most familiar annotation tasks involve identifying theoretical concepts. This includes noting the topic of the text, the position of the actor or the tone of the text. Crowdcoding seems suitable for this task (Van Atteveldt, Van der Velden, and Boukes 2021). Yet, having only a few workers annotating the majority of texts has raised concerns about data diversity and models' ability to generalize beyond the crowd-workers. In a series of experiments, (Geva, Goldberg, and Berant 2019) show that often models do not generalize well to annotations from annotators that did not contribute to the training set, suggesting that annotator bias should be monitored during data set creation. One such potential bias, especially in times of increasing polarization (Iyengar et al. 2019; Gidron, Adams, and Horne 2019; Boxell, Gentzkow, and Shapiro 2022), is based on the ideological position of the annotator. Given that annotators on crowd-coding platforms tend to be younger and hold more liberal values than the general public (Clifford, Jewell, and Waggoner 2015; Huff and Tingley 2015), this could potentially hamper the data diversity and generalizability of the model. An additional reason to monitor the annotators' ideological position as a potential source of annotator bias is that a recent study in NLP showed that experiential factors influence the consistency of how political ideologies are perceived (Q. Shen and Rose 2021). Their finding challenges the "ground-truth" assumption we as researchers make that a position for example is either left-leaning or right-leaning. People with different ideological backgrounds might experience that position differently. This challenges our way of data collection: When we are interested in the effect of e.g. elite communication, we often allow

for heterogeneous treatment effects in experimental work. This indicates that we do not assume that the treatment, most commonly using text, has the same effect for different partisans. Yet, at the same time, we forget or ignore that knowledge when creating large data-sets for our machine learning models.

The field of Natural Language Processing also works on automatically classifying texts on labels of concepts such as stance, sentiment, and political orientation. These models are trained on data created by human annotators, including a final step where disagreements and differences in annotations are leveled by aggregating, averaging, or other ways of coming to a consensus on one label for one example – i.e. a "ground truth" (Aroyo and Welty 2015b). Differences from this ground truth label are seen as errors that need to be removed or sorted out. Models then learn to predict labels for new examples based on this ground truth. Recently, there is some discussion on how realistic it is to have just one label for each example, especially for subjective or complex concepts and/or texts with multiple interpretations.[2] Annotation procedures for classification models run into myths such as "disagreement is bad" and "one annotation is enough" (Aroyo and Welty 2015b). Ambiguity and linguistic complexity should be considered for disagreement in annotations (Plank, Hovy, and Søgaard 2014): Not all linguistic examples are created equal. Moreover, disagreement can be informative for the researcher. It can for instance be used to validate hypotheses about how universal the perceptions of such concepts are (Sommerauer, Fokkens, and Vossen 2020).

Additionally, doubts on smoothing out disagreement in annotation have focused on the lack of diversity. When only annotating with one label or annotator, we get homogeneity, which is undesirable in subjective and social tasks (Geva, Goldberg, and Berant 2019) such as hate speech detection or political affiliation classification. *Whose perspective is being recorded in these datasets, and then later in the models trained on these datasets?* Framing arbitrary representations in data as "bias" misses the political character of data sets: There is no neutral data and no apolitical standpoint from which we can call out bias. Data sets are

[2]Disagreement even occurs in seemingly objective tasks such as Part of Speech tagging (Fornaciari et al. 2021; Plank, Hovy, and Søgaard 2014).

always "a worldview" and, as such, data always remains biased" (Miceli, Posada, and Yang 2022, 5). The answers of the annotators in turn influence how machine learning examples classify new models. For instance, hate speech and abuse detection are NLP tasks where race and gender of annotators influences both annotators and model performance (Gordon et al. 2022; Larimore et al. 2021; Waseem 2016). Language is inherently connected to society and culture: J. H. Shen et al. (n.d.) analyze sentiment analysis, and find that human annotators lead to certain perspectives on sentiment being recorded; notably African American English dialects are often misunderstood by such models.

Most recently, new annotation paradigms have gone one step further by asking whether we are modeling the task, or the annotator (Geva, Goldberg, and Berant 2019). Because the logical coherence task Natural Language Inference showed that annotators have several valid interpretations that are not reflected in one ground truth label, Pavlick and Kwiatkowski (2019) call for new training paradigms reflecting "the full range of possible human inferences" (p.688). Recent approaches in NLP have sought to explicitly incorporate disagreement and diversity in training data annotations, such an explicitly subjective annotation paradigm (Röttger et al. 2022), "perspectivism" (Cabitza, Campagner, and Basile 2023), and "jury learning" (Gordon et al. 2022). Political ideology is not an inherent concept in many texts, but rather dependent on who is asked to annotate and their perceptions and background (Q. Shen and Rose 2021). Extra-linguistic factors, such as annotators' own political ideology and also knowledge, influenced annotation and in turn model performance (Thorn Jakobsen et al. 2022). To test whether people with different ideological backgrounds experience a political position differently, challenging our ground-truth assumption in data annotation, we propose the following hypothesis:

> **Ideological Bias hypothesis** (*H1a*): The larger the ideological distance between respondent and the party, the less likely respondents annotate statements according to the party's uttered position.

A complex concept such as political ideology is much more likely to lead to multiple

interpretations. We call such sentences with more possible interpretations and less explicit standpoints "underspecified". A lack of explicitness in the annotated text is one of the main causes of the disagreements in earlier literature. Thorn Jakobsen et al. (2022) deduce that annotator bias comes from a process known as the affect heuristic (Slovic et al. 2007): Making a decision based on the emotional response related to your own personal attitude towards the discussed topic, especially when the text is relatively ambiguous. We therefore expect that the strength of H1a is conditional on this ambiguity:

> **Ideological Bias hypothesis** (*H1b*): The effect of H1a is stronger for underspecified sentences.

In the tradition of more strict interpretations of what constitutes as a stance especially in (computational) linguistics, sentences that are underspecified should always be annotated as "not a stance". However, this might not be the desired annotation for other research purposes: When for example the main goal is to understand how readers are affected by statements shown in e.g., a newspaper article, a more lenient definition of stance that allows annotators to infer the direction of a political stance from context and prior knowledge even though a statement is strictly speaking underspecified might be more useful. In everyday life, people will not follow strict linguistic definitions, for understanding media effects we might thus be more interested in whether stances, giving context information, are "correctly overinterpreted". To test whether people with different ideological backgrounds as well as their political knowledge might experience underidentified position differently, challenging our ground-truth assumption in data annotation, we propose the following hypotheses:

> **Ideological Overinterpretation hypothesis** (*H2a*): The larger the ideological distance between respondent and the party, the more likely respondents interpret underspecified sentences as stance.

> **Political Knowledge Overinterpretation hypothesis** (*H2b*): The more political knowledge, the more likely people interpret underspecified sentences as

stance.

In a next step, we ask: *If these biases exist, how can we alleviate them?* There have been several previous approaches to solve biased annotations, especially where it concerns political or societal aspects. Geva, Goldberg, and Berant (2019) introduce an approach where training set annotators are separated from annotators annotating data sets that evaluate the models, to ensure the evaluation is not simply accurate at replicating the original annotators, but can generalize to new annotators' judgements. However, these approaches are not aimed at reducing biases during the training set annotation procedure. Other approaches are aimed at leveraging multiple perspectives - but this is not useful when one wants one label to learn from. For the Austrian National Elections, Ennser-Jedenastik and Meyer (2018) already demonstrated that showing party labels impact annotators' assessment of the party position. So, one solution is masking. We therefore test whether masking reduces potential differentiation incited by ideological position or political knowledge with the following hypotheses:

> **Masking Solution hypothesis** (*H3a*): Masking reduces the effect of respondents' ideological position for coding stances according to the party's position.

> **Masking Solution hypothesis** (*H3b*): Masking reduces the effect of respondents' level of political knowledge for coding stances according to the party's position.

Data, Methods & Measurement

Data

We have conducted the survey experiments in the Netherlands in May 2022 and in the United States in January 2023. Both samples, recruited through KiesKompas and Prolific for respectively the Dutch and American case, consist of 3,000 participants (based on the

power analysis presented in our online compendium, and Online Appendix (OA) A-1) of 18 years and older. Both survey companies works with non-random opt-in respondents. Therefore, we measured many demographic background variables, and balance checks have been conducted to demonstrate whether certain categories are over represented in a certain experimental group (see OA A-3). Our study has been approved by the Research Ethics Review Committee of the researchers' institution. To ensure good quality of our data, one attention check (discussed in more detail in OA A-2) is included (Berinsky, Margolis, and Sances 2014).

Measurement

Experimental Conditions. Respondents are randomly assigned to either view a political party as an actor, or a masked condition, where they see `X` as an actor; simultaneously, respondents see either a fully specified sentence or an underspecified sentence, in which one needs additional information outside of the text to determine an actor's position. Table 1 gives an overview of the variations in treatment in the surveys.

Dependent Variable. We rely on whether or not a party's (implied) stance is coded according to the party's position (H1 and H3) as well as whether or not the statement is coded as a stance at all (H3). For each issue, we ask the respondent `what is according to the sentence above the position of [ACTOR]?`, with the answer categories: `in favor`, `against`, `no stance`, `don't know`. We use both a very strict interpretation of stance – specification of change and direction – and a lenient interpretation – specification of change. Using the strict interpretation, respondents are correct if they say `no stance` for the underspecified sentences and `against`, `in favor`, `in favor`, and `against` to the specified sentences one to four. Using a more lenient interpretation, respondents could say either `in favor` or `against` as well for underspecified sentences two to four.

Moderating Covariates. *Ideological position* is measured using an 11-point scale ranging from left (`0`) to right (`10`). *Political knowledge* is measured with six items from the Dutch

Table 1: Survey Questions - Experimental Conditions

Condition	US Experiment	NL Experiment
Specified	[Republicans/X] say immigration should be made more difficult.	[PVV/X] says immigration should be made harder.
Specified	[Democrats/X] say we need to put a tax on carbon emissions	[GreenLeft/X] says nitrogen emissions need to be reduced.
Specified	[Democrats/X] say we should implement a wealth tax for the richest Americans.	[Labour Party/X] says tax rate should go up for highest earners.
Specified	[Republicans/X] say the U.S. needs to consider military build-up in the Pacific Ocean	[Forum for Democracy/X] says that membership in the European Union has been especially bad for the Netherlands so far.
Underspecified	[Republicans/ X] say many immigrants are crossing our borders.	[PVV/X] says many immigrants are coming this way.
Underspecified	[Democrats/X] say carbon emissions policy should be implemented differently.	[GreenLeft/X] says nitrogen policy must be different.
Underspecified	[Democrats/X] say the tax system should be implemented differently.	[Labour Party/X] says tax system must be changed.
Underspecified	[Republicans/ X] say there should be a different military presence in the Pacific Ocean.	[Forum for Democracy/X] says the Netherlands should have a different role in the European Union.

Parliamentary Election Studies for the Dutch sample, and the three items from the American National Election Studies. [3].

Control Variables. In our analysis, we control for demographic information (gender, age, education, income, religion, job) as well as political background variables (trust in politics, ideological position on economic left-right scale and cultural progressive-conservative scale, and evaluations and prospects of the economy). Tables A.5 till A.17 in the OA demonstrate the descriptive information per country.

[3]The questionnaire can be found in OA A-1

Results

Before discussing the results of our preregistered analysis, we will describe which profiles, based on ideology (see Figure 1) and political knowledge (see Figure 2). OA A-5 gives even more detail; it describes the average profile of the respondents who annotate correctly and incorrectly (where respondents who annotated some stances correctly and some incorrectly are weighted by proportion (in)correct). Figure 1 demonstrates that in both countries people positioned on the ideological left of the spectrum are more likely to correctly code the specified sentences (upper-panel) to our gold standard. In general, we see that for the specified sentences, the coders where almost always in line with our gold standard. This picture is a bit different for the underspecified sentences. In principle, all these sentences where defined so that they are linguistacally neutral (no stance). Yet, we see that all cases but the sentence "many immigrants are crossing the border" when the party is masked, a stance is interpreted. Except for the sentences "there should be a different military presence in the Pacific Ocean" and "carbon emissions should be implemented differently" when the Democrats are shown, all sentences are interpreted as negative. What is interesting is that in the underspecified sentences, which are likely to also be in political information we have annotators for, there is quite a bit of variation on how one codes it depending on their ideological position. Looking at the profiles based on political knowledge, Figure 2 demonstrates that in the US, people with lower levels of political knowledge are more likely to be correct, whereas in the Netherlands, people with higher levels of political knowledge are more likely to be correct. This indicates that biases based on political knowledge can be context dependent. For the rest, Figure 2 paints a similar picture as 1: Respondents where to high degrees correct based on the gold standard in the specified sentences, but interpreted the underspecified sentences overwhelmingly as a negative stance.

Figure 1: Profile Ideological Distance

Figure 2: Profile Political Knowledge

To answer whether there is an ideological or knowledge-based annotation bias, we have conducted a two-by-two experiment. To test our hypotheses, we will conduct a multilevel model, with respondents clustered in issues (for more details, see OA A-4). OA A-6 displays the effect of the experimental conditions on the four dependent variables – correctly identifying a stance and over-interpreting a stance for both a strict and a lenient interpretation of stance. OA A-7 demonstrates the exploratory analyses that test the robustness of our results presented below.

First, we test whether there is an ideological bias in interpreting stances (H1a), and if this bias increases for those that are further away from the ideological position of the political actor in the under-specified condition (H1b). Figure 3 demonstrates on the regression coefficients and the predicted effects. The upper-left panel of Figure 3 demonstrates the coefficient of ideological distances for the likelihood of interpreting the stance correctly. There is a negligible positive effect – a coefficient of `0.002` – that is borderline significant in the Netherlands, when looking at the strict interpretation of a stance. Substantially, this means that there is no effect of ideological distance for correctly interpreting the stance in either the American or Dutch case. Hence, no ideological bias found, thus no support for our H1a. Looking at the lower-left panel of Figure 3, we see a small but significant effect of the interaction between ideological distance and the under-specified condition for the strict interpretations in the Dutch case (hypothesized direction), and a small negative effect for the lenient interpretation of a stance in the US case (other direction as hypothesized). Those who are ideologically further away from the party are less likely to be wrong than those who are close to the party in the Netherlands, as shown in the right-hand panel of Figure 3. For the Dutch case, using the strict interpretation, the difference is about `10%` – from a coefficient of `0.3` to `0.4`. In the American case, the slope for the lenient interpretation goes down as predicted, but not that much, with a difference of about `5%`. This could indicate that in times of heightened polarization, ideological annotation bias can be a concern. We thus find mixed evidence for H1b.

Secondly, we hypothesized that there is a risk of over-interpretation from those that are ideologically distant to the political actor (H2a) as well as those who have high levels of political knowledge (H2b). We measure this with an interaction between the experimental condition of specification and the variable of interest. In both countries, we see that the different interpretations of stance have opposite effects: A strict interpretation decreases the chance of overinterpreting, and this is exacerbated by ideological distance but not by political knowledge. Using a more lenient interpretation, we see that this increases the chance of overinterpreting. While this is not further diminished by political knowledge, ideological distance does further diminish the change in the American case – i.e. against expectation H2b. In the Dutch case, however, the further you are from the party the more likely you are to overinterpret the sentence as a stance. So, while we do find some support for our H2a, we will reflect on the interpretation of a stance in combination with the context, as this effects how crowds interpret underspecified sentences.

Figure 3: Ideological Distance

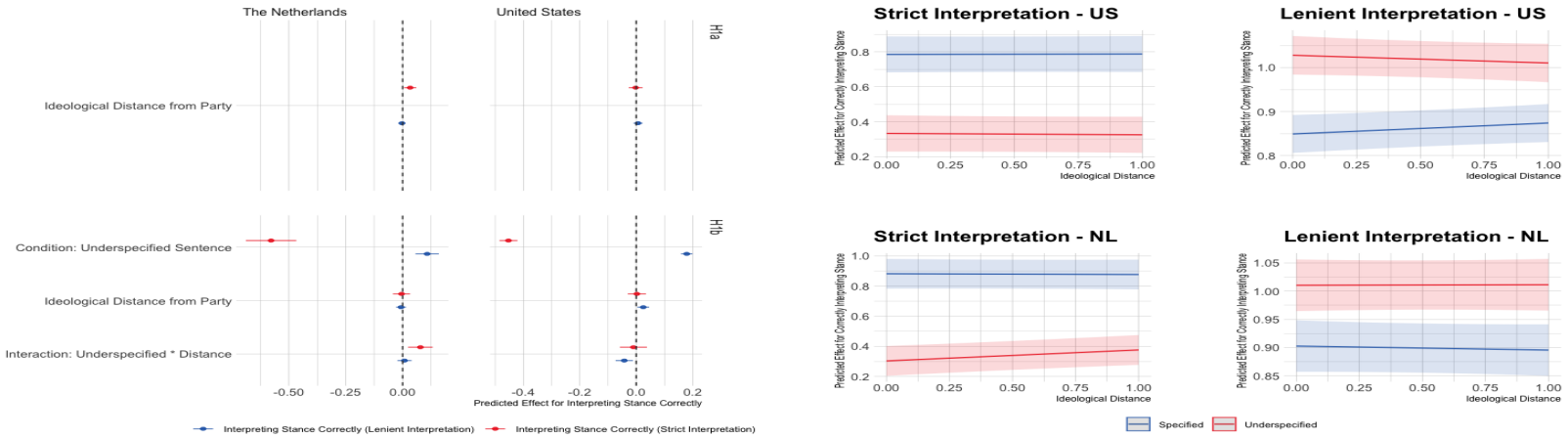
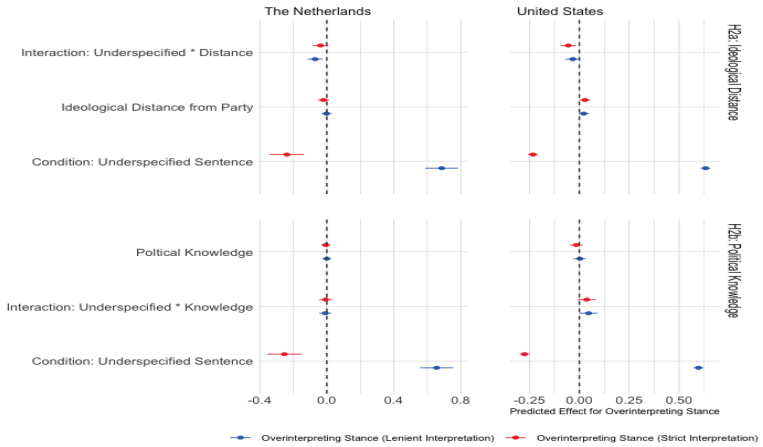
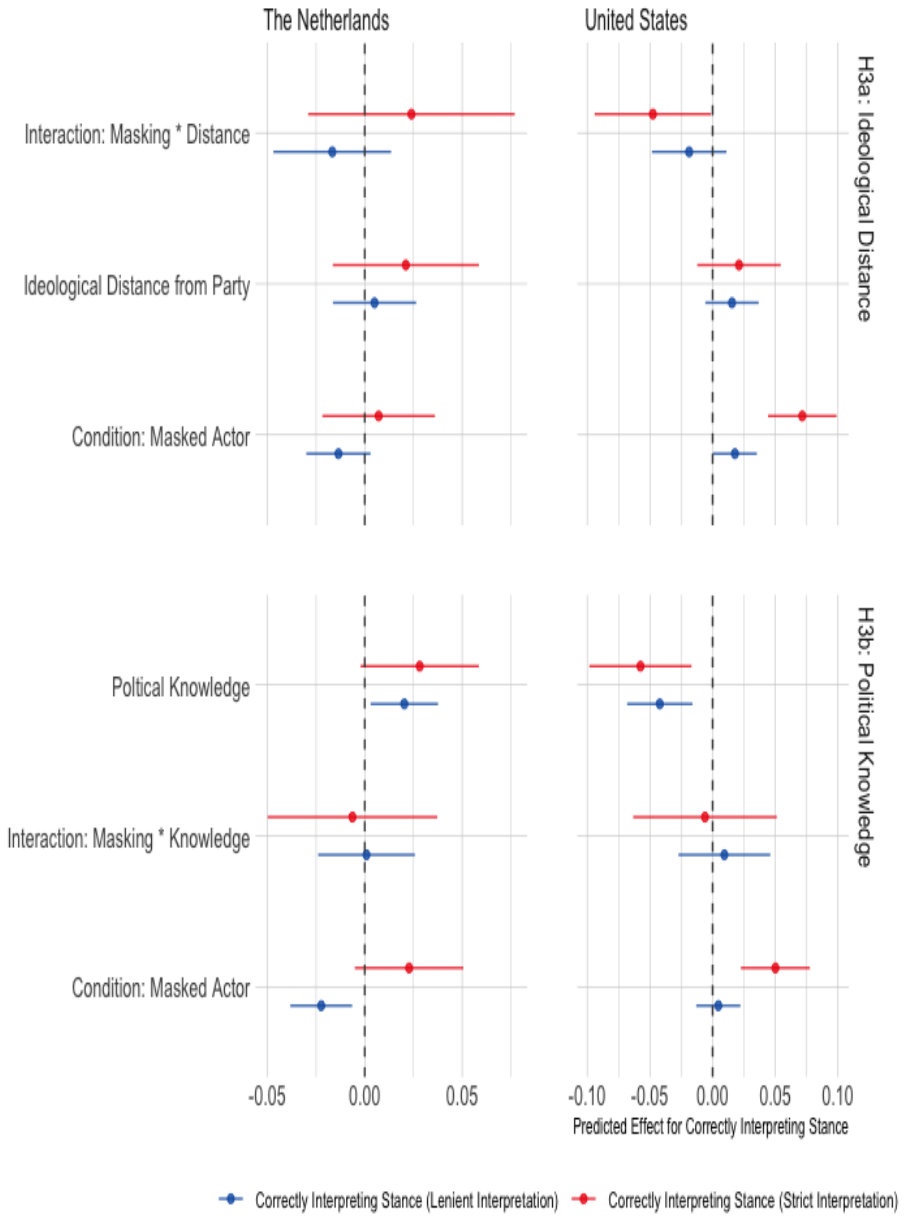


Figure 4: Results Level of Sentence Specification



Thirdly, we test whether masking of the political actor is a solution for potentially misinterpreting the stance. We hypothesized that masking should reduce the ideological bias (H3a) as well as the bias resulting from political knowledge (H3b). We test these hypotheses with an interaction between the condition masking and the variables of interest and Figure 5 demonstrates the regression coefficients. There is no effect found in the American or Dutch case regarding ideological distance. The same goes for the interaction between masking and political knowledge, on the bottom-panel of Figure 5. There is no significant effect found in the US nor int the Netherlands. This means that masking of political actors does not help to correctly interpret a sentence as a stance – i.e. no support for H3a and H3b.

Figure 5: Results Masking Solution



Discussion

Human annotation represents a crucial instrument in the toolkit of the social sciences. In order to address concerns regarding the validity and reliability of annotation tasks, we es-

tablish strict standards. Especially in times where due to the democratization of data and the advances in NLP more data can be analyzed or classified, these standards are more important than ever. After all, an algorithm trained on biased data will reproduce and often exacerbate bias (e.g. see Prost, Thain, and Bolukbasi 2019). It is key to realize that the underlying assumption to create valid and reliable annotated data is that there is only one correct interpretation for every input example, and that disagreement between the annotators is something that needs to be dealt with at all costs (Aroyo and Welty 2015a). In times of increasing factual opinion polarization (Rekker and Harteveld 2022; Lee et al. 2021), it is crucial to ask *whose perspective is being recorded in these datasets*, given that there is no neutral data and no apolitical standpoint from which we can call out bias. Our tools to create valid and reliable annotation data does currently not allow for dealing with different perspectives. In this paper, we examined annotation bias in the classification of political stances using two pre-registered crowd-coding experiments in the Netherlands and the United States. We used both a strict and a lenient interpretation of stance to illustrate whether decisions made in the research process on what constitutes a stance influences the results.

First, we tested whether ideological bias affects the interpretation of a party's stance by analyzing whether a person's ideological distance to a party affects how they classify the party's stance. Although we did not find a main effect of ideological distance (H1a), we did find evidence that the effect of ideological distance is greater for underspecified sentences (H1b) in the US, depending on how one operationalizes stances. The exploratory analyses (see OA A-7) showed that these findings where robust against different specifications of ideological distance. This conditional finding is important, because machine translation is on the rise (Licht 2023; De Vries, Schoonvelde, and Schumacher 2018): Social scientists use this technique for example to comparatively analyze political stances. Currently, most scholars using text analysis rely on English text or annotated data, translating the original source text to English (Baden et al. 2022). Our results show that the way a political stance

is defined and the context, potentially the level of polarization, affects whether annotators induced bias. Hence, naively using US annotated data and auto-translating your source text could have downstream consequences for the classification of your stances.

Second, we looked at differences in terms of overinterpretation, which we define as coding stance in an underspecified sentence (i.e. in which the stance is not explicitly made clear) using both a strict (computational linguistics) and more lenient interpretation of what a stance is. In other words, we tested when annotators inferred the stance of the party on the issue based on prior knowledge or beliefs about the party. We expected that overinterpretation is more likely if the ideological distance of the coder to the party is greater (H2a) and if the coder has more political knowledge (H2b). Overall, our findings did not support these hypotheses. Yet, they imply that ideological distance can affect how stance is processed. Regarding political knowledge (H2b), our results showed again country differences as well as that the way one defines a stance matters. For sentences with clear political stances on an issue (e.g., immigration should be made more difficult) annotators generally agreed on the stance that is expressed. But if the sentence refers to the issue without making a clear stance (e.g., many immigrants are crossing our borders) we do see a clear effect of political knowledge. This indicates that for annotation bias to occur, there has to be sufficient room for different interpretations. If the stance is clearly specified, then there is more of a ground truth, and the opinions and perspectives of the annotator therefore matter less. It is when the stance is underspecified that perspectivism (Cabitza, Campagner, and Basile 2023; Havens et al. 2022) comes into play. When a party talks about a political issue without taking an explicit stance, annotators need to fill in the gaps and make their own inferences about the party's stance on this issue. Considering that in many political texts, specifically reporting on political news, stances are not made explicit, underspecified sentences can be assumed to rather be the rule than the exception. Our results therefore imply that political knowledge is indeed a relevant dimension in shaping how an annotator perceives political stances.

Finally, we investigated whether annotation biases due to ideological distance from the party (H3a) and political knowledge (H3b) could be alleviated by masking the political party. Here we again found country differences as well as variation based on how a stance was defined. We showed that masking has no effect to resolve bias in the US context, but it does help in a less polarized context like the Netherlands. Given that the US is one of the front runners in terms of polarization, it is important to realize that this affects our annotators, and therefore annotated data, too. Biases seem to be stronger and less easy to resolve.

Our results imply that to improve annotated data for automated text analyses, and for stance detection models in particular, we need to critically evaluate how we create our gold standards. In the best case scenario, ignoring personal differences on dimensions such as political ideology only adds noise to the data. However, this is only the case if a sufficiently large sample of annotators is randomly drawn from the population. In reality, gold standards are often the product of a handful of expert coders, who typically study or work at universities. Even if a good amount of crowd coders is used, it cannot blindly be assumed that this provides a diverse and balanced reflection of political ideologies in society, as we also observed in our samples. Especially in times where an increasing amount of crowd coding answers are based on large language models such as ChatGPT which (by and large) simulate the same ideological position over and over again (Veselovsky, Ribeiro, and West 2023), reflecting on the composition and quality of annotators and how it influences gold standards becomes more than just a methodological exercise. Our results indicate that an annotation team that is skewed towards one end of the political spectrum could indeed result in a biased gold standard.

Furthermore, if the perception of political stance is contingent on ideological differences, then it can be inherently problematic to conceptualize political stance, as expressed in a text, as a ground truth. Depending on the goal of a study, it might be more appropriate to conceptualize political stance as having an inherently subjective quality, and strive to

measure this subjective space. As a general guideline, we propose making a distinction between content analysis research that makes inferences about the author of a text, and research that makes inferences about the audience. For example, a study that compares political stances between news outlets to analyze political bias in the news does require a consistent and reliable measurement of political stances. This could indeed require a ground-truth definition of political stance, that is drilled into annotators through well-defined and specific instructions and extensive training. But if the study involves making inferences about how the audience interprets political stances, for example in media effects research, then a ground-truth definition of political stance is problematic. To make accurate inferences about how a citizen subjectively perceived political content, and consequently how this might have affected behavior such as voting, we need to learn more about how shared personal characteristics relate to shared modes of perspective.

References

Akyürek, Afra Feyza, Lei Guo, Randa Elanwar, Prakash Ishwar, Margrit Betke, and Derry Tanti Wijaya. 2020. "Multi-Label and Multilingual News Framing Analysis." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8614–24. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.763.

Aroyo, Lora, and Chris Welty. 2015a. "Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation." *AI Magazine* 36 (1): 15–24. https://doi.org/10.1609/aimag.v36i1.2564.

———. 2015b. "Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation." *AI Magazine* 36 (1): 15–24. https://doi.org/10.1609/aimag.v36i1.2564.

Baden, Christian, Christian Pipal, Martijn Schoonvelde, and Mariken AC G van der Velden. 2022. "Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda." *Communication Methods and Measures* 16 (1): 1–18. https://doi.org/10.1080/19312458.2021.2015574.

Barberá, Pablo, Amber E Boydstun, Suzanna Linn, Ryan McMahon, and Jonathan Nagler. 2021. "Automated Text Classification of News Articles: A Practical Guide." *Political Analysis* 29 (1): 19–42. https://doi.org/10.1017/pan.2020.8.

Benoit, Kenneth, Drew Conway, Benjamin E Lauderdale, Michael Laver, and Slava Mikhaylov. 2016. "Crowd-Sourced Text Analysis: Reproducible and Agile Production of Political Data." *American Political Science Review* 110 (2): 278–95. https://doi.org/10.1017/s0003055416000058.

Berinsky, Adam J, Michele F Margolis, and Michael W Sances. 2014. "Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys." *American Journal of Political Science* 58 (3): 739–53. https://doi.org/10.1111/ajps.12081.

Boxell, Levi, Matthew Gentzkow, and Jesse M Shapiro. 2022. "Cross-Country Trends in Affective Polarization." *Review of Economics and Statistics*, 1–60. https://doi.org/10.1162/rest_a_01160.

Cabitza, Federico, Andrea Campagner, and Valerio Basile. 2023. "Toward a Perspectivist Turn in Ground Truthing for Predictive Computing." In *Proceedings of the AAAI Conference on Artificial Intelligence*, 37:6860–68. 6.

Clifford, Scott, Ryan M Jewell, and Philip D Waggoner. 2015. "Are Samples Drawn from Mechanical Turk Valid for Research on Political Ideology?" *Research & Politics* 2 (4): 2053168015622072. https://doi.org/10.1177/2053168015622072.

De Vries, Erik, Martijn Schoonvelde, and Gijs Schumacher. 2018. "No Longer Lost in Translation: Evidence That Google Translate Works for Comparative Bag-of-Words Text Applications." *Political Analysis* 26 (4): 417–30. https://doi.org/10.1017/pan.2018.26.

DeBell, Matthew. 2013. "Harder Than It Looks: Coding Political Knowledge on the ANES." *Political Analysis* 21 (4). https://doi.org/10.1093/pan/mpt010.

Dolinsky, Alona, Martijn Schoonvelde, Christian Pipal, Christian Baden, Fabienne Lind, Guy Shababo, Mariken A. C. G. van der Velden, and Avital Zalik. 2023. "Challenges for Multilingual Computational Text Analysis Research: Evidence from a Pre-Registered Survey of Social Science Researchers." In *4th ANNUAL COMPTEXT Conference 2022*,

1–21.

Ennser-Jedenastik, Laurenz, and Thomas M Meyer. 2018. "The Impact of Party Cues on Manual Coding of Political Texts." *Political Science Research and Methods* 6 (3): 625–33. https://doi.org/10.1017/psrm.2017.29.

Fornaciari, Tommaso, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. "NAACL-HLT 2021." In, 25912597. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.204.

Geva, Mor, Yoav Goldberg, and Jonathan Berant. 2019. "Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics.

Gidron, Noam, James Adams, and Will Horne. 2019. "Toward a Comparative Research Agenda on Affective Polarization in Mass Publics." *APSA Comparative Politics Newsletter* 29: 30–36.

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli. 2023. "ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks." *Proceedings of the National Academy of Sciences* 120 (30): e2305016120. https://doi.org/10.1073/pnas.2305016120.

Gordon, Mitchell L, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. 2022. "Jury Learning: Integrating Dissenting Voices into Machine Learning Models." In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–19.

Grimmer, Justin, Margaret E Roberts, and Brandon M Stewart. 2022. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press.

Havens, Lucy, Benjamin Bach, Melissa Terras, and Beatrice Alex. 2022. "Beyond Explanation: A Case for Exploratory Text Visualizations of Non-Aggregated, Annotated Datasets." In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, 73–82.

Huff, Connor, and Dustin Tingley. 2015. "'Who Are These People?' Evaluating the Demographic Characteristics and Political Preferences of MTurk Survey Respondents." *Research & Politics* 2 (3): 2053168015604648. https://doi.org/10.1177/2053168015604648.

Iyengar, Shanto, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J Westwood. 2019. "The Origins and Consequences of Affective Polarization in the United States." *Annual Review of Political Science* 22: 129–46. https://doi.org/10.1146/annurev-polisci-051117

Larimore, Savannah, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. 2021. "Reconsidering Annotator Disagreement about Racist Language: Noise or Signal?" In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, 81–90. https://doi.org/10.18653/v1/2021.socialnlp-1.7.

Lecheler, Sophie, and Claes H De Vreese. 2019. *News Framing Effects: Theory and Practice*. Taylor & Francis.

Lee, Nathan, Brendan Nyhan, Jason Reifler, and DJ Flynn. 2021. "More Accurate, but No Less Polarized: Comparing the Factual Beliefs of Government Officials and the Public." *British Journal of Political Science* 51 (3): 1315–22. https://doi.org/10.1017/S000712342000037X.

Licht, Hauke. 2023. "Cross-Lingual Classification of Political Texts Using Multilingual Sentence Embeddings." *Political Analysis*, 1–14. https://doi.org/10.1017/pan.2022.29.

Miceli, Milagros, Julian Posada, and Tianling Yang. 2022. "Studying up Machine Learning Data: Why Talk about Bias When We Mean Power?" *Proceedings of the ACM on Human-Computer Interaction* 6 (GROUP): 1–14. https://doi.org/10.1145/3492853.

Parmar, Mihir, Swaroop Mishra, Mor Geva, and Chitta Baral. 2023. "Don't Blame the Annotator: Bias Already Starts in the Annotation Instructions." In *17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023*, 1771–81. Association for Computational Linguistics (ACL).

Pavlick, Ellie, and Tom Kwiatkowski. 2019. "Inherent Disagreements in Human Textual Inferences." *Transactions of the Association for Computational Linguistics* 7 (0): 677–94. https://transacl.org/ojs/index.php/tacl/article/view/1780.

Piskorski, Jakub, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. "Semeval-2023 Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-Lingual Setup." In *Proceedings of the the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 2343–61.

Plank, Barbara, Dirk Hovy, and Anders Søgaard. 2014. "Learning Part-of-Speech Taggers with Inter-Annotator Agreement Loss." In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 742–51.

Prost, Flavien, Nithum Thain, and Tolga Bolukbasi. 2019. "Debiasing Embeddings for Reduced Gender Bias in Text Classification." *GeBNLP 2019* 9573: 69.

Rekker, Roderik, and Eelco Harteveld. 2022. "Understanding Factual Belief Polarization: The Role of Trust, Political Sophistication, and Affective Polarization." *Acta Politica*, 1–28. https://doi.org/10.1057/s41269-022-00265-4.

Röttger, Paul, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. "Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks." In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 175–90.

Shen, Judy Hanwen, Lauren Fratamico, Iyad Rahwan, and Alexander M Rush. n.d. "Darling or Babygirl? Investigating Stylistic Bias in Sentiment Analysis," 5.

Shen, Qinlan, and Carolyn Rose. 2021. "What Sounds 'Right' to Me? Experiential Factors in the Perception of Political Ideology." In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 1762–71. https://doi.org/10.18653/v1/2021.eacl-main.152.

Slovic, Paul, Melissa L Finucane, Ellen Peters, and Donald G MacGregor. 2007. "The Affect Heuristic." *European Journal of Operational Research* 177 (3): 1333–52.

Sommerauer, Pia. 2020. "Why Is Penguin More Similar to Polar Bear Than to Sea Gull? Analyzing Conceptual Knowledge in Distributional Models." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, 134–42. https://doi.org/10.18653/v1/2020.acl-srw.18.

Sommerauer, Pia, Antske Fokkens, and Piek Vossen. 2020. "Would You Describe a Leopard as Yellow? Evaluating Crowd-Annotations with Justified and Informative Disagreement." In *Proceedings of the 28th International Conference on Computational Linguistics*, 4798–4809.

Struthers, Cory L, Christopher Hare, and Ryan Bakker. 2020. "Bridging the Pond: Measuring Policy Positions in the United States and Europe." *Political Science Research and Methods* 8 (4): 677–91. https://doi.org/10.1017/psrm.2019.22.

Thorn Jakobsen, Terne Sasha, Maria Barrett, Anders Søgaard, and David Lassen. 2022. "LAW-LREC 2022." In, 4461. Marseille, France: European Language Resources Association. https://aclanthology.org/2022.law-1.6.

Van Aelst, Peter, and Stefaan Walgrave. 2016. "Political Agenda Setting by the Mass Media: Ten Years of Research, 2005–2015." *Handbook of Public Policy Agenda Setting*, 157–79. https://doi.org/10.4337/9781784715922.00018.

Van Atteveldt, Wouter, Damian Trilling, and Carlos Arcila Calderon. 2022. *Computational Analysis of Communication*. John Wiley & Sons.

Van Atteveldt, Wouter, Mariken ACG Van der Velden, and Mark Boukes. 2021. "The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms." *Communication Methods and Measures* 15 (2): 121–40. https://doi.org/10.1080/19312458.2020.1869198.

Veselovsky, Veniamin, Manoel Horta Ribeiro, and Robert West. 2023. "Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks." https://arxiv.org/abs/2306.07899.

Waseem, Zeerak. 2016. "Are You a Racist or Am i Seeing Things? Annotator Influence on Hate Speech Detection on Twitter." In, 138142. Austin, Texas: Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-5618.

Winter, Nicholas JG, Adam G Hughes, and Lynn M Sanders. 2020. "Online Coders, Open Codebooks: New Opportunities for Content Analysis of Political Communication." *Political Science Research and Methods* 8 (4): 731–46. https://doi.org/10.1017/psrm.2019.4.

Ying, Luwei, Jacob M Montgomery, and Brandon M Stewart. 2022. "Topics, Concepts, and Measurement: A Crowdsourced Procedure for Validating Topics as Measures." *Political Analysis* 30 (4): 570–89. https://doi.org/10.1017/pan.2021.33.