ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Rauh, Christian

Article — Published Version

Validating a sentiment dictionary for German political language—a workbench note

Journal of Information Technology & Politics

Provided in Cooperation with:
WZB Berlin Social Science Center

Suggested Citation: Rauh, Christian (2018) : Validating a sentiment dictionary for German political language—a workbench note, Journal of Information Technology & Politics, ISSN 1933-169X, Taylor & Francis, Abingdon, Vol. 15, Iss. 4, pp. 319-343, https://doi.org/10.1080/19331681.2018.1485608

This Version is available at:
https://hdl.handle.net/10419/180851

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.


https://creativecommons.org/licenses/by/4.0/

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

JOURNAL OF INFORMATION TECHNOLOGY & POLITICS
2018, VOL. 15, NO. 4, 319–343
https://doi.org/10.1080/19331681.2018.1485608

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

Validating a sentiment dictionary for German political language—a workbench note

Christian Rauh

WZB Berlin Social Science Center, Department Global Governance

ABSTRACT

Automated sentiment scoring offers relevant empirical information for many political science applications. However, apart from English language resources, validated dictionaries are rare. This note introduces a German sentiment dictionary and assesses its performance against human intuition in parliamentary speeches, party manifestos, and media coverage. The tool published with this note is indeed able to discriminate positive and negative political language. But the validation exercises indicate that positive language is easier to detect than negative language, while the scores are numerically biased to zero. This warrants caution when interpreting sentiment scores as interval or even ratio scales in applied research.

ARTICLE HISTORY
Received 14 April 2017
Revised 2 March 2018
Accepted 2 May 2018

KEYWORDS
Sentiment analysis; sentiment dictionary; text analysis; political language; German

Introduction

Most political science theories have observable implications regarding the positions, stances, and opinions voiced in spoken or written text messages. Whether societal actors communicate in a positive, neutral, or negative manner about political objects presents valuable empirical information that helps disentangling arguments about contemporary collective decision-making.

Thus, automated sentiment analyses have a high appeal for applied empirical research in political science. With the increasing digital availability of political messages, scoring large amounts of texts along predefined sentiment weights of the contained terms rests on intuitive assumptions and comes with a high level of human supervision. But this only works if the underlying sentiment weights adequately reflect term usage in the political context of interest. The technically more advanced literature recently offered context-specific machine-learning approaches (e.g., Ceron, Curini, & Iacus, 2016; Hopkins & King, 2010; Oliveira, Bermejo, & dos Santos, 2017; Van Atteveldt, Kleinnijenhuis, Ruigrok, & Schlobach, 2008), sometimes paired with crowd-sourced training data (Lehmann and Zobel 2018; Haselmayer & Jenny, 2017), in this regard. Yet, especially in projects where expressed sentiment is only one variable in a broader analytical setup, the computational, financial, or human resources required for such approaches can quickly offset the comparative advantages that led to conducting an automated analysis in the first place. In contrast, sentiment analyses based on readily available dictionaries are much less costly to implement but require information on dictionary validity. While there is a validated list of English terms (Young & Soroka, 2012), similar resources for other language contexts are rare (Mohammad, 2016).

Thus, this note provides and tests a dictionary for analyzing sentiment expressed in German political language. The employed dictionary as well as all replication materials are permanently available at https://doi.org/10.7910/DVN/BKBXWD (last accessed: 03.05.2018). When using these tools, please refer to this article as well as the SentiWS and GPC dictionaries on which it is built.

After briefly introducing the basics of dictionary-based sentiment analyses in 'Premises, promises and problems of automated sentiment scoring' section, I discuss, combine, and augment two linguistic sentiment dictionaries in the following section. The third section applies these resources in three typical settings of political language—parliamentary speeches, party manifestos, and media coverage—to assess their performance against human judgment. These tests highlight that in particular, the augmented dictionary reliably distinguished positive and negative messages

CONTACT Christian Rauh rauh@wzb.eu

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/witp.

© 2018 The Author(s). Published by Taylor & Francis Group, LLC
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

in political contexts. But they also draw attention to more general methodological issues that warrant caution for interpreting dictionary-based sentiment scores as interval or even ratio-scales. The 'Conclusions' section pulls the findings together and provides a couple of pragmatic suggestions for applied research.

Premises, promises, and problems of automated sentiment scoring

Most schools of contemporary political science would agree that politics happens in and through some form of text—be it speech acts, position papers, negotiation protocols or media reports, and commentaries, for example. Modern political science has thus quite intensively made use of content analysis methods, but the increasing availability of such texts in digital form has sparked particular interest in the (semi) automated analysis of large document corpora (for overviews, see Cardie & Wilkerson, 2008; Grimmer & Stewart, 2013). More and more, automated text analysis spills over into different subdisciplines of applied political science (e.g., Klüver, 2011; Ramey, Klingler, & Hollibaugh, 2016; Rauh & Bödeker, 2016; Wilkerson, Smith, & Stramp, 2015).

Corresponding methods often rely on rather strong assumptions about text generation or build on machine-learning techniques, but dictionary-based methods follow a much simpler intuition. They rate texts along predefined term lists referring to *a priori* known categories. Counting such term-level markers in the texts of interests then provides the basis for inferring whether a text relates to one or another of these categories. This rather simple idea has been employed since the early days of automated content analysis, with the General Inquirer (Stone, Dunphy, Smith, & Ogilvie, 1966) and DICTION (Hart, 1984) being the landmark political science examples.

Already these early applications aimed at capturing the subjectivity transported in and by political messages. The exact terminology and conceptualization vary over quasi- synonyms of 'subjectivity' and 'sentiment' including, for example, 'appraisal,' 'polarity,' 'tone,, or 'valence.'[1] What unites these approaches is the idea that the affective content of texts produced in the political process reveals information about the underlying opinions, stances, and attitudes. In this vein, large-scale quantitative information about the subjectivity in political messages is very appealing for various niches of political sciences: if texts are carefully selected to capture the political objects or actors of interest, then the sentiment expressed in these messages can be interpreted as communicated political stances that present relevant empirical evidence for an extreme breadth of political science research questions.

To measure the sentiment of political messages then, the analysis resorts to predefined lists of terms supplying quantitative weights on positive and negative connotations, counts the presence of these terms in the texts of interest, and finally aggregates their relative rate of occurrence to some sort of comparative measure, usually by normalizing it to the overall number of terms in the given text. A typical net sentiment score (cf. Young & Soroka, 2012: 215) is thus given by

$$\text{Sentiment} = \frac{\#\text{ positive terms} - \#\text{ negative terms}}{\#\text{ all terms}} \quad (1)$$

This measure, also used in the remainder of this note, is then interpreted as a relative gap between positively and negatively connoted language. In a seemingly convenient manner, it ranges between −1 and +1 where a score of .5, for example, is interpreted as 50% point overweight of positively connoted language, indicating a fairly positive sentiment of the text.

These sentiment scores appear to be a rather straightforward means to comparatively analyze subjective stances in large amounts of political messages. First, the assumption that the sentiment expressed in a piece of text is a function of the sentiment born by its individual terms seems pretty intuitive. Second, the method is rather transparent and replicable. And third, implementation is easy: Once a machine-readable dictionary with term-level sentiment weights is available, counting and aggregating is a rather trivial task for most modern data analysis environments.

[1]There is also no terminological consensus in linguistics or computer sciences. Correspondingly, what I here describe as dictionary-based sentiment analysis may also be presented as 'opinion mining,' 'subjectivity analysis,' or 'appraisal extraction' (e.g., see Pang & Lee, 2008: esp. Section 1.5).

Taken together, dictionary-based sentiment analysis seems substantially relevant, is rather intuitive, comes with a high level of human supervision, and is relatively easy to implement.

However, these premises and promises can also be questioned. While sentiment analyses are extremely reliable and transparent, simply employing an off-the-shelf dictionary does not automatically lead to valid conclusions. Seen through the lens of canonical methodological discussions about content analyses (Krippendorff, 2003: esp. Ch. 13), especially semantic and structural validity of sentiment scores are at stake.

First, sentiment dictionaries, like all content analysis tools, are invariably *context-dependent*. Consider that the term-level weights in most publically available dictionaries are actually generated in online marketing applications (Pang & Lee, 2008). But the expression of positive or negative sentiment does not have to work along identical terms in, say, a shopper's review of the most recent SLR camera on the one hand, and in the prime minister's speech on a current foreign policy crisis on the other. At best, such dictionaries contain many terms not used in political language, which leads to inefficient sentiment scores. At worst, terms in these dictionaries hold a positive connotation in their original context, while conveying a negative tone in political contexts or vice versa.

Second, from the perspective of structural validity in a given language, the assumption that sentiment in a text is a simple function of individual term weights could be an oversimplification. In this regard, *irony and negation* are key challenges. In both cases, a human receiver of the respective text message would easily spot that the sentiment of individual terms is cancelled or even flipped. A simple word-count algorithm, however, fails to do so. Irony has actually little term-level markers while negation can, in principle, be captured by going beyond the analysis of unigrams and incorporating at least some grammatical rules of the given language. In any case, political scientists would want to know how the analysis' ignorance of irony and negation affects their interpretations of sentiment.

Third, an often overlooked challenge is the *representativeness of the employed dictionary* in a given language context. Sentiment scores will arguably only be efficient and unbiased if the respective dictionary terms occur roughly as frequently in the overall language as terms with similar 'true' sentiment weights that have not made it to the dictionary. Given the power-law distributions of term frequencies (Zipf, 1935), this seems to be a pretty heroic assumption. In any case, larger dictionaries are preferable. Furthermore, dictionary-based sentiment analyses implicitly assume that a dictionary's internal balance of positive and negative terms reflects the corresponding balance in the overall language. Otherwise, sentiment scores calculated along Equation (1) cannot be interpreted as a ratio-scale, on which a score of 0 actually reflects 'neutrality.' To avoid these three validity pitfalls, in summary, an ideal sentiment dictionary would

- be as encompassing as possible,
- reflect the balance of positive to negative terms in the respective language,
- offer means to handle negated terms, and
- reflect term-level sentiment as used in political language.

The 'true' values of these criteria are unknown, otherwise a 'smaller' sentiment lexicon would not be needed. But they serve as useful guides in comparing and optimizing dictionaries. Ultimately, however, dictionary performance has to be assessed against reasonable benchmarks in a well-known environment that represents typical language usage for the envisaged applications. The remainder of this note follows these ideas. The next section discusses an optimization of existing German sentiment resources before the subsequent section compares their performance against human coders and plausible expectations in typical contexts of German political language.

Constructing a sentiment dictionary for German political language

Typically, dictionary construction starts in an inductive manner by letting humans judge the term- or document-level sentiment of example texts drawn from the context of envisaged application. Further terms are then often added by resorting to frequent co-occurrences, collocations, or synonyms of the

terms for which sentiment orientation is already known. With a view to build a resource that works across different contexts of German political language, however, I refrained from constructing the dictionary from scratch. Rather, I exploit resources that capture sentiment in the German language more generally to only then optimize the dictionary for political science applications with the above derived criteria in mind.

I start from two freely accessible and widely cited German sentiment lexicons developed by computer linguists.[2] The first resource used is the *Sentiment Wortschatz*, or *SentiWS* for short, developed at the Natural Language Processing Department of the University of Leipzig (Remus, Quasthoff, & Heyer, 2010).[3] It contains 1,650 negative and 1,818 positive words (adjectives, adverbs, nouns, and verbs) which resolve to 16,406 positive and 16,328 negative terms if the supplied inflections are taken into account. These lists were constructed along a semiautomated, three-step procedure. First, the authors automatically translated the *General Inquirer* categories 'Pos' and 'Neg' (Stone et al., 1966) and manually revised the German results. Second, they analyzed term frequencies in a set of 5,100 positively and 5,100 negatively rated online product reviews, identified the 200 most discriminating terms by a statistical co-occurrence analysis, and added them to the dictionary. Third, to retrieve additional term-level markers for positive and negative sentiment, they fed the resulting list into a German collocation dictionary to retrieve additional terms with high semantic similarity. In the original source, the resulting *SentiWS* dictionary was validated against term-level sentiment rates by two human coders in 480 sentences randomly drawn from various online fora.

The second resource used here is the *GermanPolarityClues* lexicon (*GPC*, Waltinger, 2010a, b).[4] It is also built along a multistep procedure. Starting point is the automatic translation of two existing English language dictionaries—*Subjectivity Clues* (Wiebe, Wilson, & Cardie, 2005) and *SentiSpin* (Takamura, Inui, & Okumura, 2005)—where up to three German translations were accepted for each term and sentiment direction is inherited from the English sources. The results of this translation were then manually revised to remove ambiguous terms and to enrich the dictionary further with the most positive and negative synonyms of the existing terms. The final *GPC* dictionary comes with 17,535 positive and 19,825 negative terms and was so far validated against a support vector machine classifier trained on 1,000 Amazon product reviews.

Both resources offer rather general and encompassing lists of term-level sentiment in the German language. But with regard to validity in explicitly political contexts, three caveats should be noted: First, despite their lengths, the actual content of both dictionaries is far from being identical. *SentiWS* offers four positive terms not contained in the *GPC*, but the latter offers a surplus of 2,064 positive and 4,421 negative terms. Second, only the *GPC* offers negation control with 290 bigrams to capture selected negation patterns. And third, both dictionaries were developed and so far mainly validated in the context of online product reviews while their construction also involved quite some human interpretation with unknown biases. Against the criteria for valid sentiment dictionaries derived in the preceding section, this clearly leaves room for dictionary optimization and calls for succinct testing in political language contexts.

In the quest to render the dictionary as encompassing as possible, I first combined both term sets. Then I constructed a simple regular expression to reflect bigram negations of each term in the resulting dictionary, flipping its sentiment weight.[5] To optimize

[2]After intense research based on German linguistics departments and conferences, this actually seems to be the population of German sentiment lexicons that are publically available under Creative Commons licenses. One possible addition is the 'Leipzig Affective Norms' lexicon which, however, are limited to 1,000 nouns rated into more detailed emotional categories that do not easily map onto a more general positive/negative connotation (Kanske & Kotz, 2010).

[3]See http://asv.informatik.uni-leipzig.de/download/sentiws.html (last accessed: 21.07.2016) for documentation, license, and raw data. For the paper at hand, version 1.8c of the dictionary was used.

[4]See http://www.ulliwaltinger.de/tag/german-polarity-clues/ (last accessed: 21.07.2016) for documentation, examples, and data. For the applications presented here, version 21042014 is used.

[5]Specifically, the regulation follows the pattern '(nicht|nichts|kein|keine|keinen) TERM,' thus capturing all German ways of direct bigram negation while still disregarding more complex negation patterns on the clause level, for example. The 'Dictionary validity in typical settings of German political language' section provides some orientation on how frequently these patterns affect the resulting sentiment scores.

context fit *a priori*, I opted against the fine-grained polarity weights supplied by each dictionary as these were statistically derived from online review data. Rather, I resort to the general positive (+1) or negative connotation (−1) only. Finally, the combined list was subjected to an intense manual review which, among other things, removed or corrected terms that arguably have no or an ambiguous sentiment in German political language.[6] Note, furthermore, that the analyses presented here avoid part-of-speech (POS) tagging for computational reasons. In effect, this means that terms with identical spelling but different POS tags in the dictionary have been collapsed to single observations while 11 terms for which the term-level sentiment weight differed across the POS function had to be removed.[7] Taken together, the subsequent validation exercises resort to the two extant linguistic resources and the augmented sentiment dictionary summarized in Table 1.

Some implementation intricacies have to be noted. The dictionaries are prepared such that each term comes with exactly one whitespace left and right to identify full terms only. This has to be matched in the text vectors to be scored. Furthermore, preprocessing as used in the following included the complete removal of numbers and punctuation as well as setting the remaining terms to lower case. Furthermore, the bigram negation patterns in the augmented dictionary are first replaced by a unigram marker in the text vectors before the latter are counted and aggregated into the final sentiment scores.[8] Finally, given their extreme frequency and absent sentiment information, I opted for removing stopwords before counting the overall number of terms in the texts to be scored.[9] This only affects the denominator of Equation (1), rendering our aggregate sentiment scores somewhat more evenly distributed. Along these parameters, we can now have a look on how well the dictionaries perform in settings that should be rather close to typical political science applications.

Table 1. Description of german sentiment dictionaries

	SentiWS	GermanPolarityClues	Augmented dictionary
Positive terms (Weight = 1)	15,475	17,535	17,330
Negative terms (Weight = −1)	15,404	19,825	19,750
Total terms	30,879	37,360	37,080
Positive/ Negative ratio	1.005	0.884	0.877
Negation correction	–	290 bigram phrases	Complete bigram inversion (nicht|nichts|kein|keine|keinen)

Dictionary validity in typical settings of German political language

Plenary debates: rating random sentences from Bundestag speeches

The first validation effort resorts to parliamentary debates and relies on a random sample of 1,500 sentences drawn from the corpus of all MP speeches in the *German Bundestag* 1991–2013 as presented by Rauh, De Wilde, and Schwalbach (2017).[10] The sampled sentences contain between 1 and 139 terms with a mean length of 19.26 (see Appendix A for details). Such short coding units present a very challenging context for dictionary-

[6]This first concerned 51 terms that seemed either completely irrelevant for political contexts—such as 'Schwimmschlammzerstörer' (floating sludge destructor). But it also concerned 225 terms that are presumably frequently mentioned in political language for which, however, no *a-priori* sentiment assumption seemed reasonable—consider examples such as 'Abstimmung' (coordination or vote), 'Debatte' (debate), 'Zoll' (tariff or customs), but also 'Arbeitslosigkeit' ('unemployment'). Finally, I added 145 terms representing missing inflections of terms that were already in one of the source dictionaries. A full list of manual changes is provided in the replication package accompanying this manuscript.

[7]For example, the German term 'Würde' would come with a positive sentiment as a noun where it means 'dignity,' while it has a slightly negative sentiment as a verb where it denotes conjugative speech meaning 'would.'

[8]For example, 'nicht schön' or 'keine freude' is not part of the dictionary itself but have to be replaced by 'NOT_schön' or 'NOT_freude' in the source texts and is then matched by sentiment weights of −1 accordingly.

[9]I have employed the list of 231 German stopwords used in the Snowball Stemmer project, which can be accessed at http://snowball.tartarus.org/algorithms/german/stop.txt (last accessed: 25.07.2016).

[10]The sentences were collected by randomly sampling 1.500 of all 149.832 speeches in the Bundestag corpus. Based on punctuation patterns, each speech was then sliced into individual sentences to then randomly pick one sentence from each speech. The sampling script is provided in the replication package.

based sentiment analyses, simply because there is little term-level information that can be matched by the counting algorithm. But in prominent political science content analyses—think of quasi-sentences in the Comparative Manifesto Project (CMP) (Volkens, Bara, & Budge, 2009, see also below), the core-sentence approach in mobilization studies (Hutter, Grande, & Kriesi, 2016), or claims analysis in the public sphere literature (Koopmans & Statham, 1999)—short coding units of approximately single sentence length are rather common.

So, how do the sentiment dictionaries fare in this environment? Let us first look at the distributions of scores in Figure 1. Two observations are particularly noteworthy. First, the univariate distributions on the diagonal of the figure are extremely 'peaked' with much more observations at the zero points than a normal distribution would suggest. This is most pronounced for *SentiWS* which results in a score of 0 in 666 of the 1,500 sentences (44.4%). But while its spread is visibly larger, the strong tendency toward supposedly neutral scores also holds for the considerably larger *GPC* dictionary which retrieves zero scores for 498 (33.2%) of the sampled sentences. The augmented

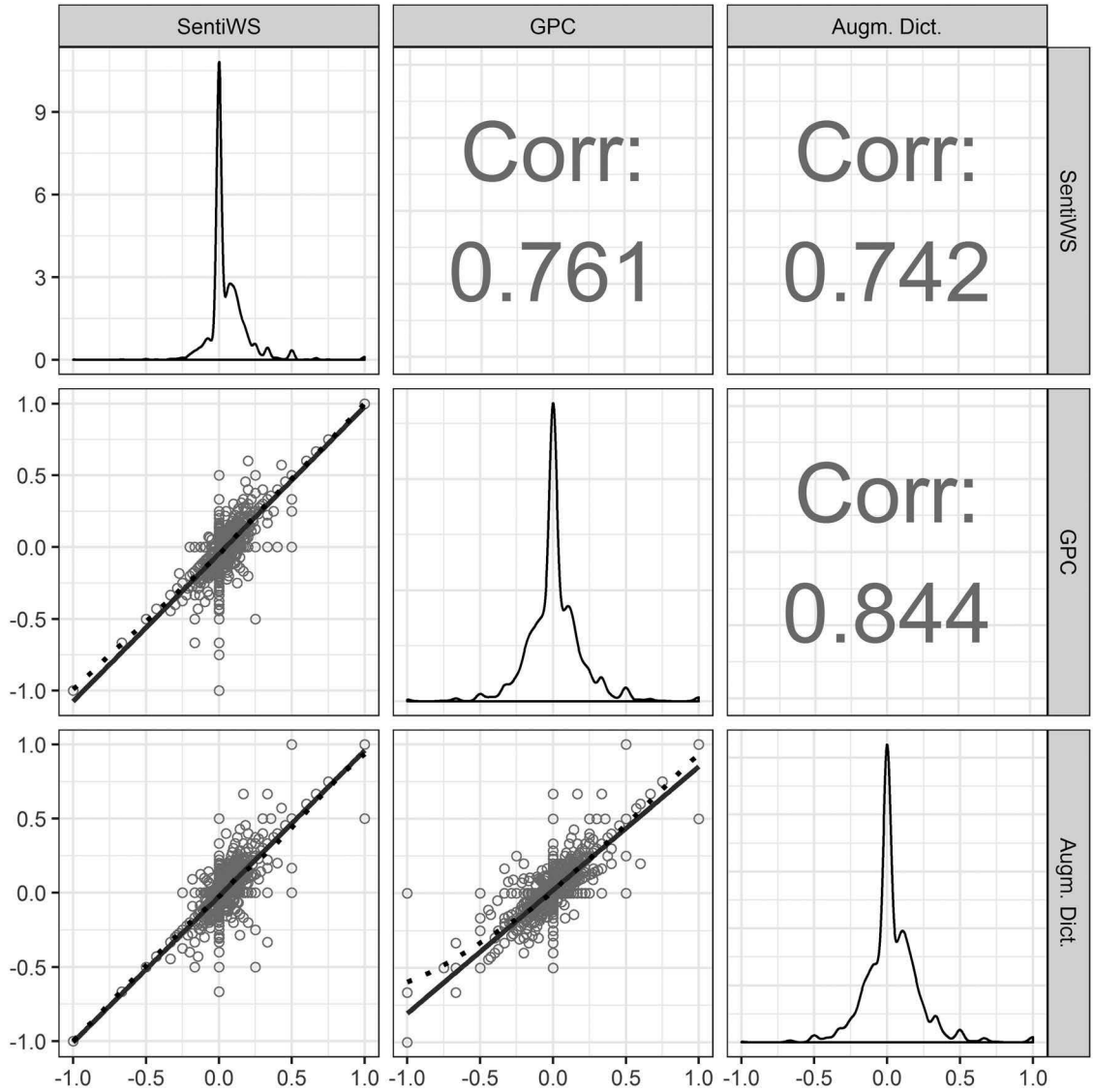

Figure 1. Univariate and bivariate distributions of different sentiment scores in a sample of 1,500 Bundestag sentences.

dictionary with its more enhanced negation control improves this on a minor scale: 474 of the sampled sentences (31.6%) are classified as neutral with regard to expressed sentiment.

The second key observation is that scores from the three dictionaries are positively and robustly correlated as the combination of the linear (solid line) and the more flexible locally smoothed fit (dashed line) in the lower panels shows. Framed positively, this finding indicates that all three seem to measure a similar construct and that even the rather small *SentiWS* dictionary already offers a pragmatic resource to tap into this concept. Framed negatively, in this challenging sample of small coding units, a significantly larger dictionary size and an enhanced negation control do little to improve the results.[11] In the sample of 1,500 sentences from plenary speech in the German Bundestag, the more advanced dictionaries change the sentiment judgment only for a few cases.

But for applied researchers, it is much more important to know whether this high level of agreement matches some sort of 'true' distribution of sentiment in the sentences under analysis. The obvious approach to address this question is systematically comparing the sentiment scores to the judgment of human coders. This also makes sense from an inferential point of view: Most content analyses in the social and especially the political sciences assume some sort of 'sender-message-audience' framework where the analysis of the message is not a purpose in itself but mainly serves to draw conclusions about the sender, the audience, or their relationship (Neuendorf, 2001: Ch. 3). In the running example of Bundestag speeches, for example, we might be interested in the position an MP wants to communicate to her electorate with the respective speech. Comparing automated sentiment scores to the human decoding of the respective message presents the most reasonable benchmark.

However, human judgment should not be taken as a gold standard. In fact, the methodological content analysis literature revolves almost entirely around controlling human biases, predispositions, and situational disturbances. Automated sentiment scoring should have reliability advantages in this regard. And they deliver information at a much higher resolution. Calculating the net sentiment score as presented in the preceding section results in a fine-grained scale that would overburden any comparative human judgment on sentiment. This implies that any comparison between automated scores and human judgment cannot be matched exactly and will come with some amount of error. Ideally, though, this error should not lead to systematic biases between automated and human judgment.

With this in mind, I asked three human coders to rate the 1,500 sampled sentences. The coders were one female and two male political science students, all German native speakers with no prior experience in automated text analyses and no specific information about the experiment's purposes. They accessed the individual sentences in an author-written, browser-based survey tool without further context (see screenshots in Appendix A). It asked coders for the basic sentiment that they perceive in a given sentence, allowing them to choose among 'negative,' 'positive,' or 'neutral' (with the latter option being the fallback for possibly lazy coders). Each coder rated all 1,500 sentences. In order to avoid fatigue, the coders could freely distribute the coding task along six chunks of 250 sentences each over a time span of 3 days. Running order of the sentences was randomized within and across these chunks to avoid possible halo effects.

This setting forces human coders to rely to the same term-level information that the scoring algorithms use as well. But for our purposes, important differences still apply. First, human judgment might come with personal biases triggered (e.g., by specific political issues appearing in the sentences). Second, unlike the sentiment algorithm, human coders knew that they were dealing with political language and can be assumed to adjust their judgment given prior experience in the German context. And third, we can reasonably assume that the human coders have an advantage over the algorithms in processing negation and irony correctly. In these respects, the experiment

[11]The negation control along the bigram inversion described in 'Constructing a sentiment dictionary for German political language' section affected 78 sentences (i.e., 0.05% of the sample used in this section).

isolates the typical challenges of semantic and contextual validity of dictionary-based analyses we have identified earlier.

For an initial comparison, I forced the continuous sentiment scores to the same three-point scale that the humans faced. A sentence with a sentiment score with a 1 standard deviation above the sample mean was coded as 'positive,' scores undercutting a 1 standard deviation below the mean were coded as 'negative,' and everything in between was considered 'neutral.' Based on this common scale, we treat the three humans and the three dictionaries as independent coders (detailed results in Appendix A).

This exercise shows that the three human coders themselves were not particularly well equipped to capture the sentiment transported by the individual sentences. They fully agreed in less than 60% of all cases only and with a value of .68 Krippendorff's alpha, a common reliability measure which corrects for chance agreement, is at the lower border of the area that is usually considered sufficient. This is hardly surprising given the little contextual and term-level information that was available for coding.

More important for the validation goals of this article, this initial exercise highlights lacking agreement between humans and machine-generated sentiment scores. On the three-point scale constructed here, all six coding approaches (three humans and three dictionaries) agree only in one quarter of the 1,500 sentences. The rank correlations in Appendix A exhibit a low match between humans and all three dictionaries, where the augmented dictionary performs slightly better than the *GPC* and especially the *SentiWS* resources.

Where does this disagreement between human and machine coding come from? One possible source of error is the arbitrarily chosen cutoff point for the sentiment scores used in this exercise. While the data would allow us to inductively estimate more optimal thresholds, these would hardly be generalizable to other applications. But Table 2 reveals a more general message on why humans and dictionary-based sentiment scores disagree in this sample. In fact, in most deviating cases, humans detected some sentiment while the dictionary-based scores indicated neutrality.

Table 2. Shares of 'neutral' cases in 1,500 random bundestag sentences by coder

Coder	Human 1	Human 2	Human 3	SentiWS	GPC	Augment. D.
Neutral cases	799	631	535	1,253	1,183	1,165
Share of sample	53.3%	42.0%	35.7%	83.5%	78.8%	77.7%

Note: For automated coding, a sentence is coded as 'neutral' when falling in between one sample standard deviation below and above the sample mean.

On the one hand, personal biases might push human coders to see political sentiment where there objectively is none. On the other hand, and more important for our purposes, the sentiment scoring algorithms come with an *inbuilt neutrality bias*. To see that, recall that the *GPC* dictionary, as the longest one in the current comparison, delivers 37,360 terms in total. Consider furthermore that the overall corpus of Bundestag speeches from which the classified sentences were drawn contains approximately 600,925 unique terms.[12] Assuming that this is a good representation of the German language and assuming that term-frequency distributions are normally distributed (which they tend to be not), this would mean that for every encountered term there is roughly only a probability of .06 that it is matched in the GPC dictionary. In other words, for any given term there is a 94% chance that the dictionary finds no sentiment. Taking into account that the sampled sentences are on average 19 terms long and again assuming a normal distribution of term frequencies for simplicity, the likelihood that an individual sentence contains no term-level sentiment information thus amounts to approximately $.94^{19} = .31$. Thus, our baseline expectation is to find numerical 'neutrality' in at least one third of coding units of this particular sample. And this does not yet consider that sentiment weights of individual terms might cancel each other out. The take-away message is that even fairly large sentiment dictionaries are biased toward a zero score that can hardly be always interpreted as true 'neutrality' of the text message.

[12]Against heavy debates among linguists, I use a very crude measure here. Information on unique terms is derived by pooling all speeches, removing all grammatical punctuation and numbers, setting everything to lowercase, splitting tokens along whitespaces ('\\s+'), collapsing the data to unique observations, and counting the number of resulting rows.

But this comparison is also very strict thus far: We have asked how well a sentiment score replicates human classification in individual coding units. However, we have also seen that human classification itself is far from being perfect. In many large-N applications applied research might be satisfied with a sentiment score that gets the tendency right. Thus I devise a second strategy to render human judgment and machine scoring comparable. Rather than reducing the resolution of the sentiment scores, it renders human judgment more fine grained. The idea is that ambiguities in human judgment capture the strength of the sentiment transported by a given text (cf. Young & Soroka, 2012: 215). Along this line, human judgment was classified as 'clearly positive/negative' if all three coders agreed, was coded as 'rather positive/negative' where only a majority of two coders opted for the respective sentiment, and all remaining sentences received a 'neutral' label (the latter being the case in 44.7% of the sampled sentences). Figure 2 plots the mean sentiment scores and their bootstrapped 95% confidence intervals across these five categories of human judgment.

Comparing the performance of the three dictionaries against each other, we see that the *SentiWS* on average produces positive scores only. This suggests utmost caution when interpreting sentiment scores on a ratio scale. In Table 1, we have seen that the *SentiWS* comes with a roughly equal balance of positively and negatively connoted terms, whereas the *GPC* and the augmented dictionary are tilted roughly toward a 12% overweight of negative terms each. Figure 3 suggests that the latter is closer to the 'true' balance of positive and negative terms in German political language. The scores from both the augmented dictionary and the *GPC* exhibit much steeper slopes from the most negative to the most positive categories thus equally outperforming the *SentiWS* in discriminating among human majority votes on sentiment in individual sentences. The GPC and the augmented dictionary perform almost equally well here, where the latter exhibits only a marginally wider spread between the most extreme categories of human judgment.

It should be noted, however, that this works better in the realm of positively connoted language. For both the *GPC* and the augmented dictionary, the mean differences across the 'neutral,' 'rather positive,' and 'clearly positive' are statistically significant. This does not equally hold for sentences that human coders have judged as expressing negative sentiment. While a statistical approach would discriminate the 'clearly negative' and the 'neutral' category, the 'rather negative' would not be significantly distinguishable from these extremes.

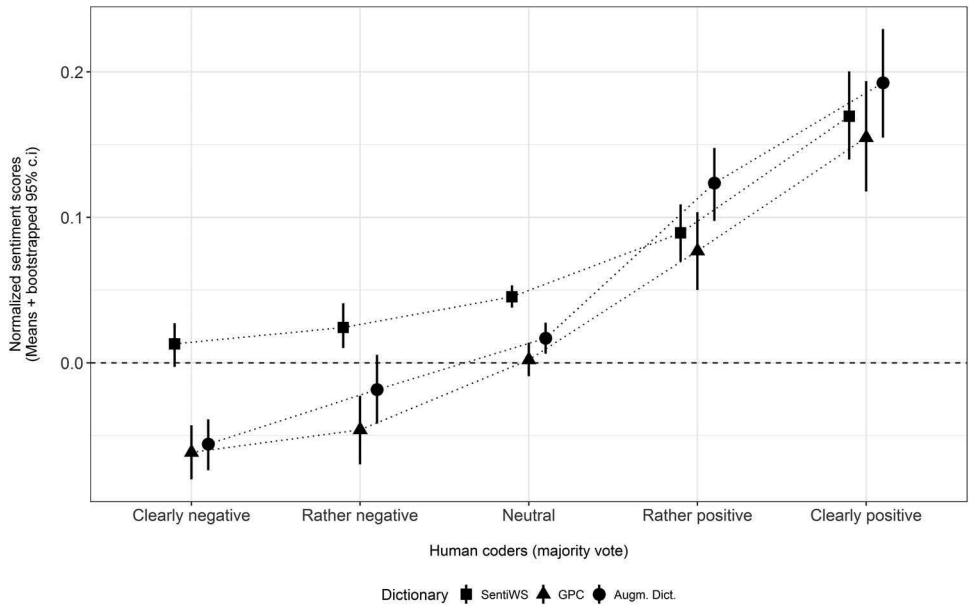

Figure 2. Human judgmentjudgment against mean sentiment scores in 1,500 random Bundestag sentences.

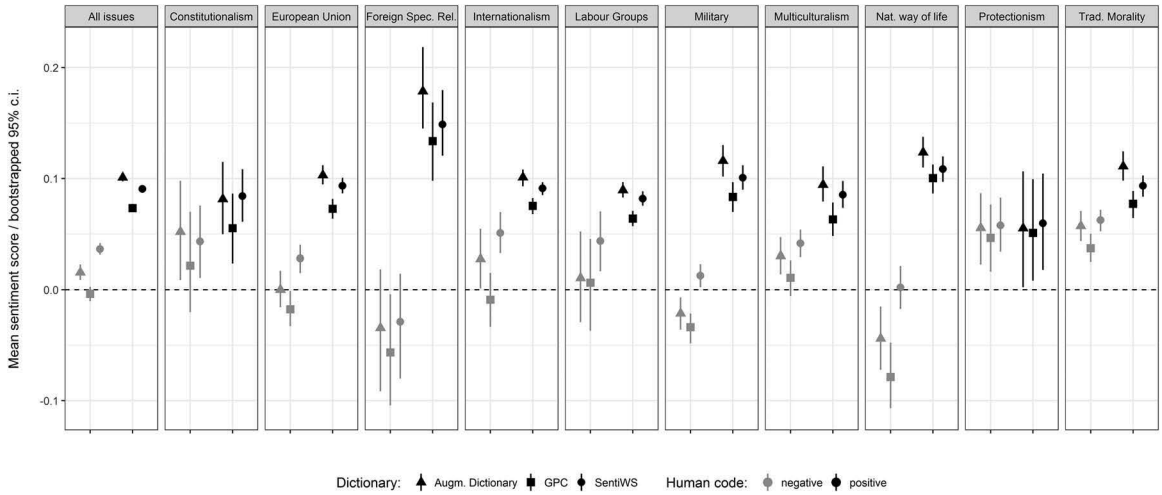

Figure 3. Comparing sentiment scores across different directional CMP categories.

A closer look on the full distribution of individual dictionary scores (Appendix A) shows that this is mainly due to a couple of 'false positives.' Thus, I qualitatively analyzed the 72 sentences on which the human coders agreed on a negative sentiment while the sentiment dictionaries indicated an overweight of positive terms. While there where a few instances that contained terms with clearly negative connotation in political contexts but no representation in the dictionary, most of the other instances could be traced to the use of subjunctive constructions, more complex negation patterns, and inferences for which some contextual knowledge about German political debates seems necessary. Table 3 presents some illustrative examples.

Complex negations and subjunctive constructions indicate natural limits of term-level sentiment coding. But the qualitative analysis indicates the presence of these phenomena is also frequently associated with a finite set of term-level markers. Examples are the ubiquitous 'nicht' for complex negation, or subjunctive verbs such as 'hätte,' 'würde,' or 'könnte.' These markers do not necessarily bear an own sentiment weight but they might provide anchor points to depress the weights of other terms in the respective context. Future research could exploit the replication data offered with this note to develop meaningful coding rules in this regard.

Furthermore, dictionary-based coding could be improved by filling missing terms. In practice, however, this is an endless endeavor that also risks tilting the balance of positive and negative terms toward the particular context from which the sample texts are drawn. For dictionary validation, the cases where human interpreters presumably used contextual knowledge to judge the sentiment of sentences from parliamentary speeches are more challenging. Contextual information—such as knowing that the otherwise positively connoted term 'beruhigen' (calm down) might indicate negative sentiment in the direct interaction of MPs—cannot be part of a dictionary- or, more generally, any bags-of-words-based method of automated text analysis. For our purposes here, contextually driven errors raise the question how well the dictionaries perform in other political contexts beyond parliamentary interaction.

Partisan campaign messages: scoring directional manifesto categories

Another context that is highly relevant for applied political science research is partisan campaign messages. Thus, this section compares dictionary performance against one of the largest political science content analyses, the CMP (see Budge, Klingemann, Volkens, Bara, & Tanenbaum, 2001; Volkens et al., 2009). The CMP relies on human coding of policy preferences expressed in electoral manifestos of now more than 1,000 parties from

Table 3. Examples of 'false-positive' sentiment scored bundestag sentences humans coded as 'clearly negative'

Sentence example	Augmented dictionary score	Source of error
Mag für eine klientelpartei wie die fdp *möglich* sein	.16	Missing term
Frau staatssekretärin, wie wollen sie denn angesichts der vorgesehenen politik des plattmachens von arbeitsplätzen eine *kompensation* in der region *ermöglichen* ?	.13	Missing term
Ich würde mir hier *deutlich klarere* vorgaben wünschen	.33	Subjunctive
Ich glaube, es ist *richtig*, dass die *interessierte* öffentlichkeit das einmütige ergebnis einer diskussion des deutschen bundestages als sehr viel *glaubwürdiger* wahrnehmen würde, als wenn vier anträge vorliegen und mehrere davon *nicht angenommen* werden, obwohl sie im grunde identisch sind	.24	Subjunctive
Ich weiß nicht, ob wir damit das *gewünschte ziel*, nichtraucher zu *schützen*, *erreichen*	.38	Complex negation
Dieser umstand ist sicher maßgeblich dafür *verantwortlich*, dass weitere ideen und vorschläge nicht mehr zum zuge kamen	.18	Complex negation
Beruhigen sie sich ein bisschen	.33	Context assumptions
Die steuern verteilen sie hin zu den vorständen der *großen* banken, *versicherungen* und angeschlossenen konzernen	.2	Context assumptions

1945 until today and publishes corresponding data in various aggregations.[13]

Human coding starts with unitizing each manifesto's text into quasi-sentences (i.e., grammatical constructs that are supposed to contain one and only one political argument). These coding units are then assigned to exactly one category from an encompassing coding scheme. While the project is firmly rooted in the saliency theory of partisan competition and thus focusses on selective issue emphasis (Budge 1982), over time some interest in positional partisan conflict became reflected in the coding scheme. Corresponding categories come as item pairs indicating the same issue with a 'positive' and a 'negative' version. For these directional categories, coders have to decide whether a quasi-sentence expresses partisan support of or opposition to a given political issue.

We will exploit these categories based on the intuition that support and opposition are traceable on the term-level as well. In other words, we assume that a party pushing an issue couches this in positively connoted language, while a party wanting to express opposition will use a much more negatively connoted terminology. Then the automatically retrieved sentiment scores should mirror the human classification into positive and negative categories. If this travels over the different political issues, we have a strong signal of the dictionaries' validity in debates with varying substantial content.

Again human classification is far from a perfect 'gold standard,' however. On the one hand, the deterministic coding and aggregation of quasi-sentences might overestimate the certainty of a data generating process where stochastic human errors affect both the writing and the parsing of manifestos (Benoit, Laver, & Mikhaylov, 2009). On the other hand, the manifesto coding efforts have also been criticized for their unobservable error. Each manifesto is usually handled by one coder only, making the results susceptible to all sorts of personal biases that adversely affect the codes' reliability (Mikhaylov, Laver, & Benoit, 2012). Thus also in this example, perfect replication of individual human codes cannot be our benchmark. Rather, the question is how well the sentiment scores get at least the tendency in the directional manifesto items right.

To build a corresponding data set, I first identified all directional categories in the manifesto scheme. Then I then accessed the most recent version of the Manifesto Corpus (Lehmann, Matthieß, Merz, Regel, & Werner, 2016) via the extremely convenient R wrapper for the database's API (Merz, Regel, & Lewandowski, 2016). The tool does not only offer the quantitative data set but also grants direct access to an extensive set of human-annotated full-text data. I retrieved all German-language quasi-sentences from German, Austrian, and Swiss party manifestos published after 1998. This leaves a total of 14,008 quasi-sentences from 26 unique parties in 11 elections between September 1998 and September 2013 available for analysis.

[13]See https://manifestoproject.wzb.eu/ (last accessed: 25.07.2016).

Table 4 shows that text availability varies over political issues. The categories 'Internationalism,' 'European Union,' and 'Labor Groups' also exhibit a strong overweight of positively connoted arguments. Furthermore with 9.6 terms on average, the individual quasi-sentences are very short. Thus automated coding might be even more strongly biased to zero. Yet, *SentiWS* scores indicate a zero in 5,505 cases (39.3%), the *GPC* in 4,423 cases (31.6%) and the augmented dictionary in only 3,929 cases (28.1%). This is not significantly worse than what we have observed in the Bundestag sentence sample above. Language in party manifestos seems more pointed, using more subjective language than parliamentary debates. Yet and still, about one third of coding units are classified as containing zero sentiment which bears heavily on our comparison here, because human coders had no 'neutral' category available.

Figure 3 plots the mean sentiment scores with bootstrapped 95% confidence intervals for the three dictionaries and the directional CMP categories available. The leftmost panel in this figure assesses the scores across all 14,008 quasi-sentences under analysis. Three things stand out. First, all three dictionaries can, on average, distinguish quasi-sentences that human coders have considered negative from those they have perceived as positive. The differences in means are statistically highly significant for each dictionary, but their discriminatory power varies: the mean difference between human-coded positive and negative categories is .054 for the *SentiWS* scores, but increases to .077 for the *GPC* and .085 for the augmented dictionary. Second, the mean sentiment scores again populate only a very limited range of the theoretically possible scale, which is arguably due to the high share of numerically 'neutral' cases highlighted earlier. And third, we again see that also the scores for categories human coders considered negative tend to land in the positive range of the scale. Like for the parliamentary sentences above, in particular the *SentiWS* dictionary but also the other two seem to perform worse in identifying negative as opposed to positive messages. In sum, the aggregate message is pretty similar to what we found above: the sentiment scores get the gist of human coding right but the findings raise doubts with regard to interpreting them on an interval and especially on a ratio scale.

The remainder of the figure highlights notable variation over the different issue categories. With regard to 'Foreign Special Relationships' and a 'National Way of Live,' the retrieved scores actually tend to be in the negative range where human coders saw opposition to these political issues. In these two categories, also the discriminatory power of the sentiment scores seems to be greatest as judged by the mean distances. And while all three dictionaries show significant mean differences in 8 of the 10 categories, neither dictionary can reasonably discriminate opposing from supportive signals in the realms of 'Constitutionalism' and 'Protectionism.'

To gain a better impression of the underlying errors, I again took a qualitative look on cases of strong disagreement between humans and machines. Table 5 presents the most extreme cases by including observations where the sentiment score from the augmented dictionary was above an absolute level of .75 in one direction and the human coder classified the quasi-sentence in exactly the opposite direction. Along this extreme criterion, the dictionary produced two 'false positives' and seven 'false negatives.'

Two things become clear. First, all of these examples are extremely short. And in all of these instances the human coders has, whether rightly or wrongly, applied quite some contextual knowledge to fit these small bits of text into the coding scheme. That the terms 'soziale Ausgrenzung' (social exclusion) and

Table 4. Available german quasi-sentences in directional CMP categories

Political issue	Human codes	CMP code	# Quasi-sentences	Average length of unit
Foreign special relationships	Positive	101	96	10.58
	Negative	102	10	11.40
Military	Positive	104	719	9.15
	Negative	105	732	9.15
Internationalism	Positive	107	2,598	9.70
	Negative	109	278	9.39
European Comm./ Union	Positive	108	1,657	9.14
	Negative	110	408	10.30
Constitutionalism	Positive	203	135	8.12
	Negative	204	51	11.27
Protectionism	Positive	406	49	10.39
	Negative	407	134	10.03
National Way of Life	Positive	601	837	8.47
	Negative	602	160	9.68
Traditional Morality	Positive	603	777	8.84
	Negative	604	624	9.58
Multiculturalism	Positive	607	666	8.86
	Negative	608	501	8.35
Labor Groups	Positive	701	3,479	8.35
	Negative	702	97	10.96

Table 5. Contradictory code examples from CMP quasi-sentences

Issue	Human code	Quasi-sentence	Augmented dictionary score
Internationalism	Negative	Unabhängige	1
Protectionism	Negative	Dafür schafft die globalisierung enorme chancen	0.8
Internationalism	Positive	Die *straflosigkeit* von menschenrechtsverbrechen *muss beendet* werden	−.75
Multiculturalism	Positive	Auslöser für *angst* und *intoleranz* ist *unwissenheit*	−1
Multiculturalism	Positive	*Soziale ausgrenzung*	−1
Multiculturalism	Positive	*Armut*	−1
Labor groups	Positive	*Geringfügig beschäftigte*	−1
Labor groups	Positive	Viele *leiden* unter *stress* und *erschöpfung*	−1

'Armut' (poverty) in examples five and six each indicate a partisan argument in support of multiculturalism can hardly be inferred from the text bits alone. Likewise seeing why the extremely positive term 'independent' in the first example is a message against internationalism is impossible without knowing the immediate context of this coding unit. Second, human coders following the CMP scheme often interpret opposition to something as support for another thing. While the sentiment scoring just sees a negative message if 'suffering,' 'stress,' and 'exhaustion' are mentioned as in example 8, the human coder—rightly or wrongly—adds workers to the picture to then interpret this as partisan support of organized labor. Similarly, from the fact a party manifesto presents globalization in extremely positive language as in example 2, the human coder has derived a signal against protectionism. Taken together, the far-off cases seem to result from a combination of strong dependency on contextual human interpretation on the one hand and very little term-level information for sentiment scoring on the other. The former is a challenge for the manifesto scheme; the latter presents a 'natural' limitation of the automated methods discussed here.

But again most applied political science research will hardly be interested in classifying short, individual quasi-sentences. Rather, the aggregate message a party sends on a given political issue in a given election will be of interest. This opens the opportunity to pool text bits into larger coding units that offer more term-level information to the scoring algorithms. Inferring a party's EU position is a typical application in this vein. Research relying on CMP data usually resorts to the so-called net-support measure which captures the overweight of pro-EU (per108) over anti-EU (per110) quasi-sentences as a share of the total number of coded units in a manifesto. The measure is not uncontested and comes with a range of implicit assumptions (Marks, Hooghe, Steenbergen, & Bakker, 2007), but given its prevalence in EU studies it presents an interesting validation benchmark. I pooled all text units coded as either pro- or anti-EU to party/election level. This reduces the data set to 56 unique observations. I then scored the text blocks with the augmented dictionary. Figure 4 plots the results against the standard human net-support measure.

Note first that the left Figure panel suggests much less neutrality bias than we have observed in the above examples. After stopword removal, the pooled EU sentences have an average length of 348.7 terms. In these longer coding units, a zero score is detected only twice by the *SentiWS* and the *GPC* dictionaries (3.6%) and only once by the augmented dictionary (1.8%). The longer the text units are, the more likely it is that a sentiment algorithm takes a decision.

The assessment of whether automated scoring replicates the typical human coding of partisan EU support repeats four messages that we have already distilled in the earlier experiments. First, the right-hand panel of Figure 8 underlines once more that the sentiment scores discriminate well between manifestos that send a supportive EU message and those that contain an overweight of EU opposition according to the human-coded net-support scale. Second, uncertainty of the mean scores is again much higher for messages that humans have coded as negative. Third, the human indicator of neutrality—the net-support measure is usually interpreted on ratio level—does hardly fall together with the zero point of the scale emerging from the automated scores. Rather, the numerical range of sentiment scores again tends toward the positive side of the theoretically possible scale. Errors in detecting negative language seem systematic and are most likely due to complex negation and subjunctive constructions. Fourth, despite getting the tendency right, there is only a weak correlation of

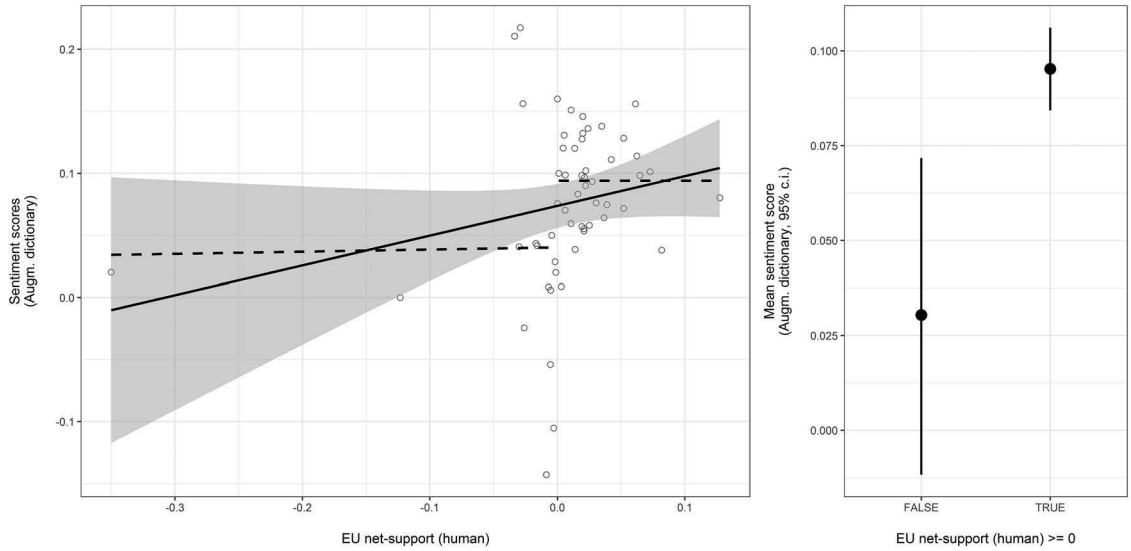

Figure 4. Sentiment scores of EU-related manifesto text vs. the human-coded net EU support measure.

.22 between the sentiment scores and the continuous net-support measure if we take the full scales of human and machine coding into account (blue regression line). This correlation seems to be almost completely leveraged by the mean differences left and right of the human scale's zero point. Looking either into the negative (left dashed regression line) or the positive range (right dashed line) of the human-coded scale only, the evidence suggests no linear relationship.

Taking these four messages together and assuming, for a moment, that the human-coded net-support scale is a valid measure of partisan EU positions, leads to one conclusion: while the augmented dictionary reliably distinguishes positive from negative messages, caution is warranted when interpreting the sentiment scores as an interval and particularly as a ratio scale.

Media coverage: scoring published sentiment during the 'Causa Wulff'

The third and final validation exercise completes the picture by analyzing dictionary performance in media coverage, another highly important context for applied political science. Rather than only comparing automated scoring to explicit human coding, this validation exercise also assesses how well the dictionary scores replicate human intuition on media reporting during a recent scandal on the German president.

The President of Germany is a rather ceremonial office awarded to persons with high moral integrity. However, on December 13, 2011, the major German tabloid *Bild* broke the news that the then President Christian Wulff had not been honest about a house loan from a befriended entrepreneur during an earlier parliamentary investigation, subjecting him to strong allegations of favoritism and unethical behavior. The affair heated up further when it became known in early January 2012 that Wulff had personally tried to stop the initial report, among other things by leaving a rant and threats on the personal mailbox of *Bild's* chief editor. After intense public debates and the initiation of a formal investigation, Wulff finally resigned on February 17, 2012, quoting the decline of publical trust as a reason. This was then followed by further revelations and debates on whether Wulff was entitled to the former Presidents' honorarium.

We exploit this 'Causa Wulff' as a natural experiment on expressed sentiment in German political news coverage. Standing above the woes of electoral competition, the nature of the office initially lets us expect neutral or even positive sentiment in news coverage on the President. This, however, should significantly change over the unfolding scandal where intuition would let us expect that the initial *Bild* report on the loan affairs and probably also the publication of the call affair are relevant break points that push expressed media sentiment on President Wulff into the negative domain. To test this expectation, I resort to the Nexis 'ZEITNG' group resource which contains full-text coverage of 265 German language newspapers, journals, and agency reports.

I retrieved all articles that featured 'Wulff' in the headline and that were published between October 2011 and April 2012 to then preprocess and score these texts as specified above. In total 4,038 individual articles matched the above criteria, with clear publication peaks after the initial report on the loan affair, the publication of the call affair and finally the resignation (Appendix C).

These articles offer a more natural and contextually complete coding unit of political text as compared to the random speech sentences and the manifesto quasi-sentences assessed above. In addition, with a sample average of 512 terms per article, they offer more term-level information available for scoring. Accordingly, we find smoother distributions and much less numerical 'neutrality.' Along the *SentiWS* only two articles receive a zero score, while the *GPC* and the augmented dictionary with their more adequate term balance and their enhanced negation controls classify 148 (3.7% of the sampled observations) and 227 cases (5.6%) as neutral, respectively. These differences warrant a closer look on disagreements among the three scoring algorithms in the sample at hand. Figure 5 plots the respective distributions. Again we see that the *SentiWS* scores fall almost entirely in the positive range of the scale and result in a rather right-skewed distribution. The *GPC* scores make the most negative judgment on average but come with an almost symmetrical distribution that, however, exhibits a minor bump close to the presumably 'neutral' zero point. Finally, with its more encompassing negation control, the augmented dictionary produces scores that distribute rather symmetrically around a mean that also comes closest to zero. In this sample, thus, the latter dictionary most closely reflects the idea of term-level net sentiment scores described in 'Premises, promises and problems of automated sentiment scoring' section.

Figure 6 furthermore highlights that the three scoring algorithms are much less correlated than in the earlier examples. With more term-level information available, the markedly longer dictionaries are much more sensitive and thus disagree more often with the *SentiWS* scores, in particular where the latter detects relatively extreme positive sentiment. But also the *GPC* and the *augmented dictionary* scores disagree to some extent on these longer coding units. The enhanced negation control of the augmented dictionary again leads to fewer cases around a value of zero sentiment and it pushes the mean upward, especially in cases where the GPC sees rather strong negative or positive sentiment. These improvements are marginal, but the distribution of the augmented dictionary scores comes closest to the idea of interpreting sentiment scores on a ratio scale.

This finding is also emphasized by comparing the automated dictionary scores of the newspaper articles to the judgment of human readers. For an initial validation in this context, I drew a random sample of 100 articles from the 'Causa Wulff' corpus and asked two additional human coders to assess whether the individual newspaper articles convey a positive or negative message.[14] The human coders agreed on the basic direction of the respective media reports in 87% of cases in this sample and like in the above Bundestag example I took their majority vote as the indicator for human judgment. 'Clearly positive' or 'clearly negative' are those articles on which both coders agreed while those with differing human judgment are considered 'neutral' or unclear. Figure 6 presents the comparison of this indicator to the mean sentiment scores derived from the three dictionaries.

Again the *SentiWS* scores only fall into the positive range of the scale and, more importantly, cannot reliably distinguish among human judgments despite more term-level information in this corpus. In contrast, the GPC and the augmented dictionary scores perform much better by exhibiting a clearly positive slope across the scale of human judgment which discriminates 'clearly negative' from 'clearly positive' judgment in a statistically significant manner despite a rather small sample size. As suggested above, the augmented dictionary approximates a ratio scale of sentiment best: The observed slope is rather linear and, unlike for the GPC, the average sentiment score for texts that humans perceive as 'clearly negative' or 'clearly positive' is significantly different from zero.

Yet, does this rather good approximation of human judgment also translate to a valid mapping of how the German media covered the political scandal during the 'Causa Wulff'? To tackle this final validation

[14]The coders have not taken part in the earlier presented experiments and were two male native German speakers with a political science background and no prior information about the purpose of the test. Article order was randomized for each coder. The sampling procedure, the resulting sample of texts as well as the human codes are available in the replication package.

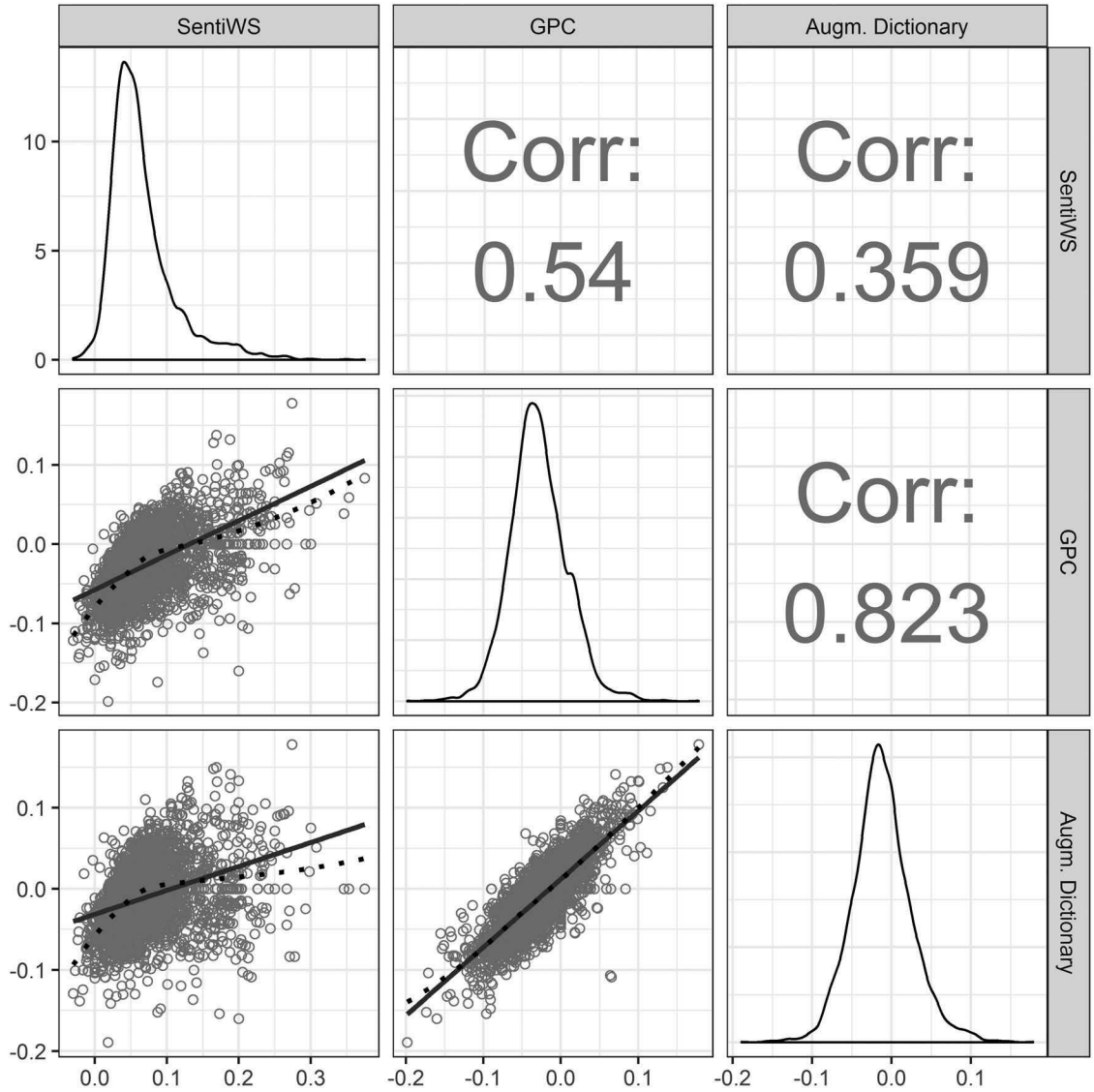

Figure 5. Univariate and bivariate distributions of different sentiment scores in a sample of 4,038 Wulff-related newspaper articles.

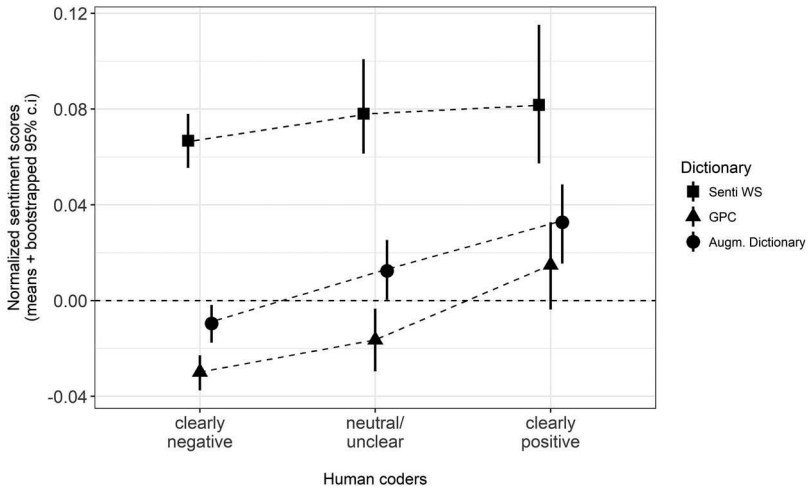

Figure 6. Automated sentiment scores vs. human coders in a random sample of 100 newspaper articles from the 'Causa Wulff' corpus.

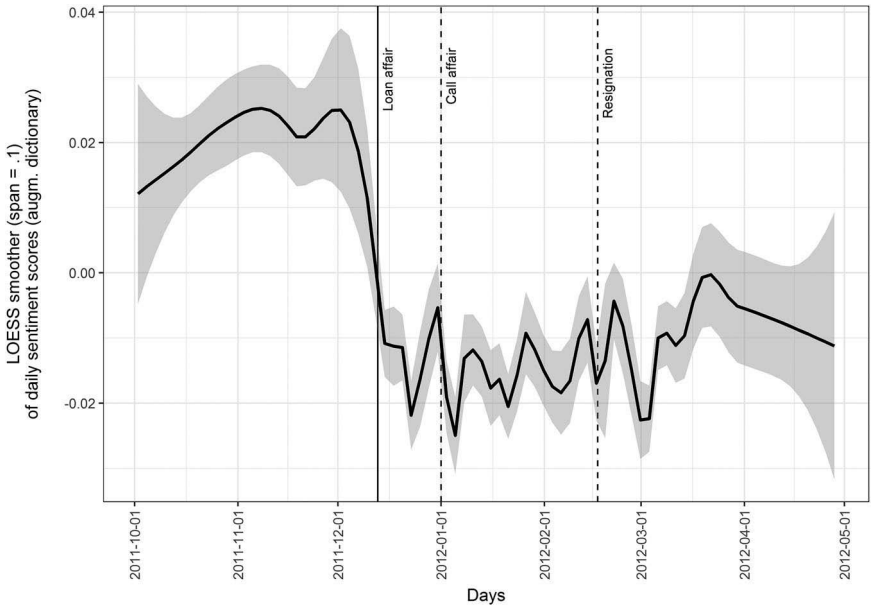

Figure 7. Sentiment in Wulff-related media coverage over the unfolding political scandal.

exercise, Figure 7 presents a slightly smoothed perspective on daily sentiment variation according to the augmented dictionary over key stages of the unfolding scandal. As expected, especially the initial publication of the loan affair in mid-December is associated with a pronounced, statistically highly significant, and lasting slump of sentiment expressed in Wulff-related media coverage. Likewise, the publication of the call affair in January is related to a statistically significant drop in sentiment in newspaper articles discussing Christian Wulff. Yet, this effect is less pronounced and temporally much more contained. In the subsequent weeks, average sentiment fluctuates but drops again shortly before Wulff's resignation when public prosecutors announced an official investigation. After the resignation, the sentiment scores briefly indicate recovery but drop again markedly in late February when the discussion on Wulff's honorarium sets in and further perks he enjoyed during private holidays become publically known.

While we lack a benchmark to validate the absolute size of these effects, the relative temporal dynamics of the dictionary-based scores clearly conform to our expectations on media sentiment during this recent scandal on the President of Germany.[15]

Conclusions

Many contemporary political theories have observable implications in textual data, which are nowadays often digitally accessible. Dictionary-based sentiment scoring thus has a high appeal for applied political science research: it promises to uncover relevant information about expressed political stances from large text corpora along a seemingly intuitive idea and at rather low cost. Adding to the availability of validated resources in this regard, this note introduces a sentiment dictionary for German political language by combining and augmenting two extant resources from computer linguistics. Three experiments in typical political science contexts—plenary speech, party manifestoes, and media coverage—produce valuable insights on the validity of the three tested resources. While the *SentiWS* dictionary does not perform very well in

[15]A brief comparison with the other dictionaries in Appendix C (Figure C3) shows that also *SentiWS* and *GPC* would identify the loan and the call affairs as significant break points in media sentiment on Wulff. But we find disagreement with regard to the size of these effects, the baseline sentiment, and the speed with which media sentiment recovers after Wulff's resignation.

political language contexts, the *GPC* lexicon already offers acceptable results when compared against human judgment and intuition. With its enhanced negation control and the removal of politically ambiguous terms, the *augmented dictionary* improves this further: like the *GPC* it reliably discriminates political language that humans receive as either positively or negatively connoted, but exhibits slightly more discriminatory power, less neutrality bias and, in result, better distributional properties As such, the augmented dictionary published with this note offers a useful addition to the toolkit available to applied political science research.

Beyond the particular resource, however, the three validation exercises also offer more general advice for applied research. Dictionary-based sentiment analysis and its standard output of normalized term overweight are not as intuitive as it seems at first sight. Three things, in particular, have to be taken into account when interpreting numerical sentiment scores in a political science context.

First, even for long dictionaries and particularly in short coding units, sentiment scores are most likely biased toward zero. This bias can be roughly assessed by comparing the dictionary size to the number of unique terms in the application's language and applying the resulting term-level likelihood of finding any sentiment to the average length of the coding unit in the sample under analysis. If the resulting baseline expectation seems inacceptable, and if theory as well as sample size allow, pooling the text of coding units is a promising remedy for this challenge.

Second, a sentiment score of or close to 0 cannot be readily interpreted as representing true neutrality of the message. The comparisons in this note highlight substantial disagreement on the absolute level of sentiment across dictionaries where the augmented dictionary most closely approximates a symmetrical distribution around the zero point. Thus, caution is warranted when interpreting sentiment scores as a ratio-scale. As a remedy, the experiments so far suggest that the best cutoff point for robust discrimination between positive and negative language is the respective sample mean. In other words, the validation exercises suggest to derive relative rather than absolute inferences from the sentiment scores. This also holds when comparing across contexts with deviating language patterns, such as the comparison across different political issues presented in the manifesto example. In such settings, within-context normalization seems warranted.

Third, the evidence presented in this note suggests that sentiment scoring works better for positive than for negative messages. This might vary over languages, but here it is apparently driven by the more complex negation and subjunctive patterns in the German language. This unequal performance, however, raises doubts on whether the resulting sentiment scores can be interpreted as an interval scale. Applied researchers are well advised to check the distributional assumptions when using sentiment in analyses requiring interval-level variables. If they are not met, reducing the scores to ordinal scales and check the finding's robustness against different cutoffs seems to be a pragmatic way forward.

Acknowledgments

I appreciate the research assistance of Felix Große-Kreul, Maximilian Lobbes, Rebecca Majewski, Johannes Scherzinger, and Xaver Keller. The manuscript has furthermore benefitted from the comments of two anonymous reviewers and from the participants and convenors of the 'Quantitative Text Analysis in Manifesto Research' workshop in July 2016 at the WZB Berlin Social Science Center, in particular the discussion by Daniel Bischof.

References

Benoit, K., Laver, M., & Mikhaylov, S. (2009). Treating words as data with error: Uncertainty in text statements of policy positions. *American Journal of Political Science*, *53*(2), 495–513. doi:10.1111/ajps.2009.53.issue-2

Budge, I. (1982). Electoral volatility: Issue effects and basic change in 23 post-war democracies. *Electoral Studies*, *1*(2), 147–168. doi:10.1016/0261-3794(82)90001-4

Budge, I., Klingemann, H.-D., Volkens, A., Bara, J., & Tanenbaum, E. (2001). *Mapping policy preferences: Estimates for parties, electors, and governments 1945–1998*. Oxford, UK: Oxford University Press.

Cardie, C., & Wilkerson, J. (2008). Text annotation for political science research. *Journal of Information Technology & Politics*, *5*(1), 1–6. doi:10.1080/19331680802149590

Ceron, A., Curini, L., & Iacus, S. M. (2016). iSA: A fast, scalable and accurate algorithm for sentiment analysis of social media content. *Information Sciences*, *367-368*, 105–124. doi:10.1016/j.ins.2016.05.052

Grimmer, J., & Stewart, B. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297. doi:10.1093/pan/mps028

Hart, R. (1984). *Verbal style and the presidency: A computer-based analysis*. Orlando, Florida, USA: Academic Press.

Haselmayer, M., & Jenny, M. (2017). Sentiment analysis of political communication: Combining a dictionary approach with crowdcoding. *Quality & Quantity* 51(6): 2623–2646.

Hopkins, D., & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, *54*(1), 229–247. doi:10.1111/j.1540-5907.2009.00428.x

Hutter, S., Grande, E., & Kriesi, H. (eds). (2016). *Politicising Europe: Integration and mass politics*. Cambridge, MA: Cambridge University Press.

Kanske, P., & Kotz, S. A. (2010). Leipzig affective norms for German: A reliability study. *Behavior Research Methods*, *42* (4), 987–991. doi:10.3758/BRM.42.4.987

Klüver, H. (2011). The contextual nature of lobbying: Explaining lobbying success in the European Union. *European Union Politics*, *12*(4), 483–506. doi:10.1177/1465116511413163

Koopmans, R., & Statham, P. (1999). Political claims analysis: Integrating protest event and political discourse approaches. *Mobilization: An International Quarterly*, *4* (2), 203–221.

Krippendorff, K. (2003). *Content analysis: An introduction to its methodology*. London, UK: Sage Publications, Inc.

Lehmann, P., Matthieß, T., Merz, N., Regel, S., & Werner, A. (2016). *Manifesto Corpus. Version 2016-3*. Berlin, Germany: WZB Berlin Social Science Center.

Lehmann, P., & Zobel, M. (2018) 'Positions and saliency of immigration in party manifestos. A novel data set using crowd coding.', European Journal of Political Research: Early View.

Marks, G., Hooghe, L., Steenbergen, M. R., & Bakker, R. (2007). Crossvalidating data on party positioning on European integration. *Electoral Studies*, *26*(1), 23–38. doi:10.1016/j.electstud.2006.03.007

Merz, N., Regel, S., & Lewandowski, J. (2016). The Manifesto Corpus: A new resource for research on political parties and quantitative text analysis. *Research & Politics*, *3*(2), 2053168016643346. doi:10.1177/2053168016643346

Mikhaylov, S., Laver, M., & Benoit, K. (2012). Coder reliability and misclassification in the human coding of party manifestos. *Political Analysis*, *20*(1), 78–91. doi:10.1093/pan/mpr047

Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In H. Meiselman (ed.), *Emotion measurement*. Amsterdam, The Netherlands: Elsevier.

Neuendorf, K. (2001). *The content analysis guidebook*. London, UK: SAGE Publications, Inc.

Oliveira, D. J. S., Bermejo, P. H. D. S., & dos Santos, P. A. (2017). Can social media reveal the preferences of voters? A comparison between sentiment analysis and traditional opinion polls. *Journal of Information Technology & Politics*, *14*(1), 34–45. doi:10.1080/19331681.2016.1214094

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1–2), 1–135. doi:10.1561/1500000011

Ramey, A., Klingler, J., & Hollibaugh, G. J. (2016). Measuring elite personality using speech. *Political Science Research and Methods: First View*.

Rauh, C., & Bödeker, S. (2016). Internationale Organisationen in der deutschen Öffentlichkeit - ein Text Mining Ansatz. In M. Lemke & G. Wiedemann (ed.), *Text-Mining in den Sozialwissenschaften. Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse*. Wiesbaden, Germany: Springer VS.

Rauh, C., De Wilde, P., & Schwalbach, J. (2017). The ParlSpeech data set: Annotated full-text vectors of 3.9 million plenary speeches in the key legislative chambers of seven European states. *Harvard Dataverse*. doi:10.7910/DVN/E4RSP9

Remus, R., Quasthoff, U., & Heyer, G. (2010) 'SentiWS – a publicly available German-language resource for sentiment analysis', *Proceedings of the 7th International Language Resources and Evaluation* LREC'10:1168–1171.

Stone, P., Dunphy, D., Smith, M., & Ogilvie, D. (1966). *General inquirer: A computer approach to content analysis*. Boston, MA: The MIT Press.

Takamura, H., Inui, T., & Okumura, M. (2005) 'Extracting semantic orientations of words using spin model', *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*: 133–140.

Van Atteveldt, W., Kleinnijenhuis, J., Ruigrok, N., & Schlobach, S. (2008). Good news or bad news? Conducting sentiment analysis on Dutch text to distinguish between positive and negative relations. *Journal of Information Technology & Politics*, *5*(1), 73–94. doi:10.1080/19331680802154145

Volkens, A., Bara, J., & Budge, I. (2009). Data quality in content analysis. The case of the Comparative Manifestos Project. *Historical Social Research/Historische Sozialforschung*, *34*(1 (127)), 234–251.

Waltinger, U. (2010a) 'GermanPolarityClues: A lexical resource for German sentiment analysis', *International Conference on Language Resources and Evaluation*, 17-23 May *2010* LREC'10.

Waltinger, U. (2010b) Sentiment analysis reloaded: A comparative study on sentiment polarity identification combining machine learning and subjectivity features, *Proceedings of the 6th International Conference on Web Information Systems and Technologies*, April 7-10,*2010*, Valencia, Spain, WEBIST '10.

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, *39*(2–3), 165–210. doi:10.1007/s10579-005-7880-9

Wilkerson, J., Smith, D., & Stramp, N. (2015). Tracing the flow of policy ideas in legislatures: A text reuse approach. *American Journal of Political Science*, *59*(4), 943–956. doi:10.1111/ajps.12175

Young, L., & Soroka, S. (2012). Affective news: The automated coding of sentiment in political texts. *Political Communication*, *29*(2), 205–231. doi:10.1080/10584609.2012.671234

Zipf, G. K. (1935). *The psychobiology of language*. Oxford, UK: Houghton-Mifflin.

Appendix A—Additional information on sample and coding of Bundestag sentences

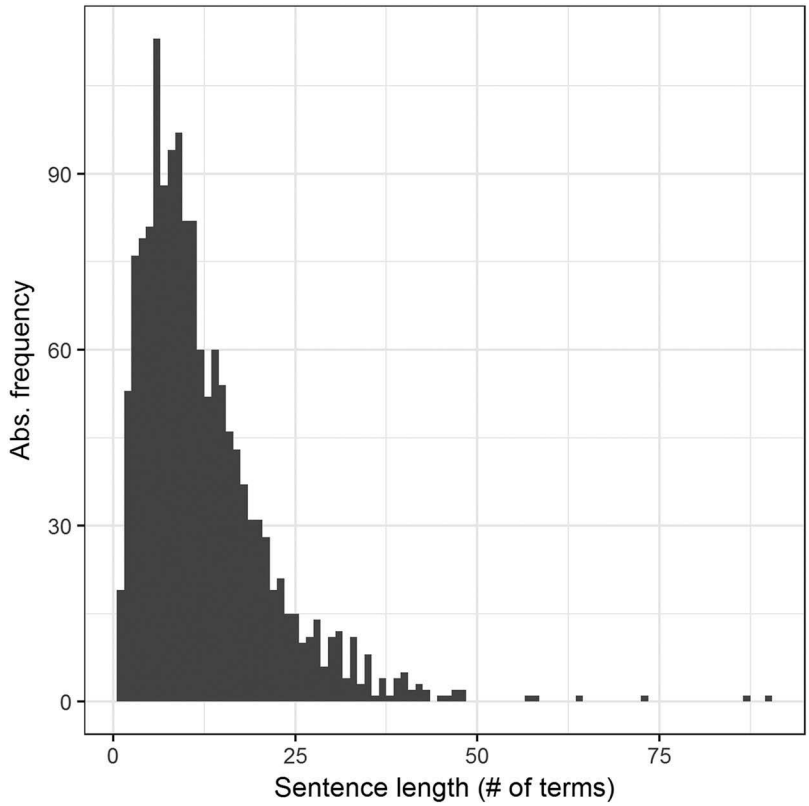

Figure A1. Comparing different dictionaries during the 'Causa Wulff.'

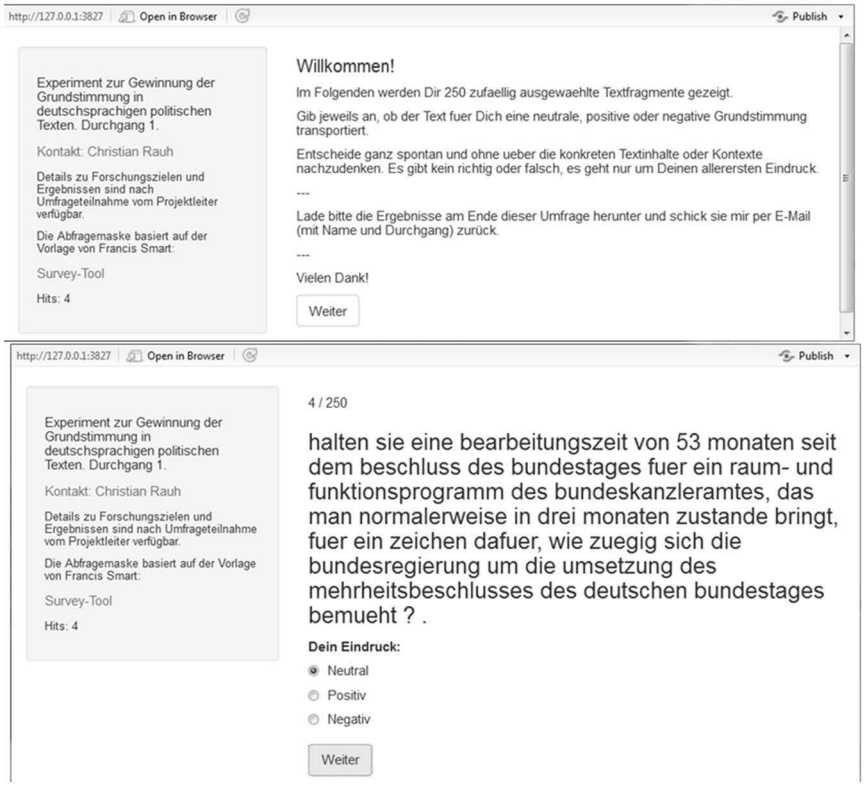

Figure A2. Length of sampled sentences.

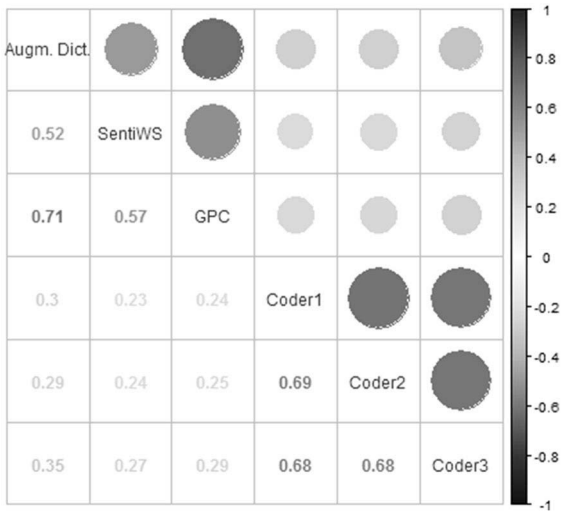

Figure A3. Screenshots of survey tool for human coders.

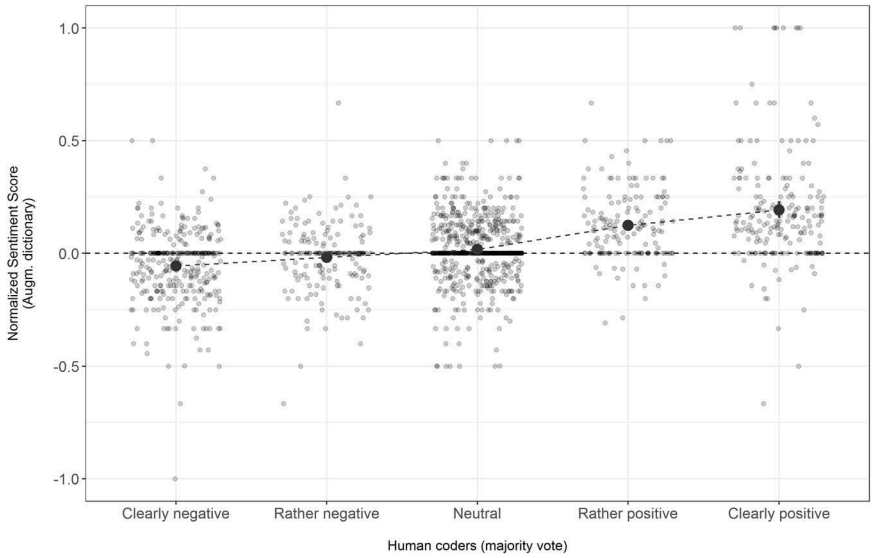

Figure A4. Rank correlations of human- and dictionary-based sentence classification.

Note: Despite their common roots also the sentiment dictionaries achieve no significantly better 'inter-coder reliability' than humans. The univariate distributions shown in Figure 1 in the main text together with the 1-standard deviation cutoff point for text classification suggest that particularly the *SentiWS* dictionary with its much lower spread disagrees frequently with the other two automated procedures. For applied research, wishing to reduce automated scores to discrete ordinal scales means that the robustness of the resulting inferences should be checked against different cutoff points.

Table A1. Inter-coder Agreement

Coders	Percentage agreement	Krippendorff α (ord. data)
Human coders (3)	59.1	.675
Sentiment dictionaries (3)	77.2	.599
Humans and dictionaries (6)	24.4	.404

Note: A total of 1,500 randomly sampled sentences from Bundestag speeches coded as 'positive,' 'neutral,' or 'negative.'

Appendix B—Additional information on scoring manifesto units

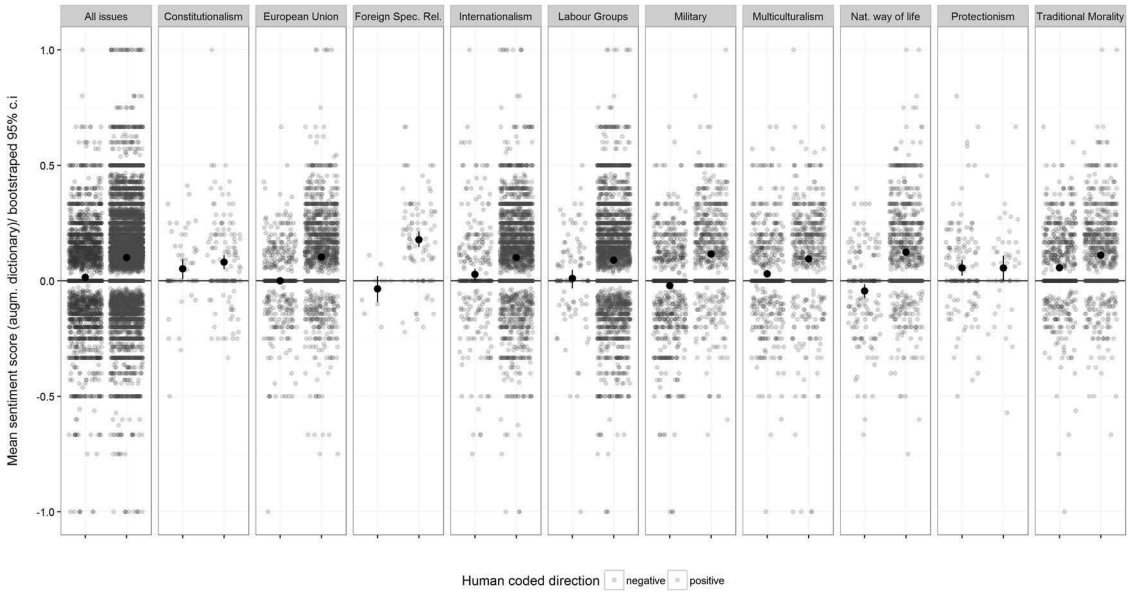

Figure B1. Distribution of augmented dictionary scores over human-coded categories.

Appendix C—Additional information on sampling/scoring 'Causa Wulff' media coverage

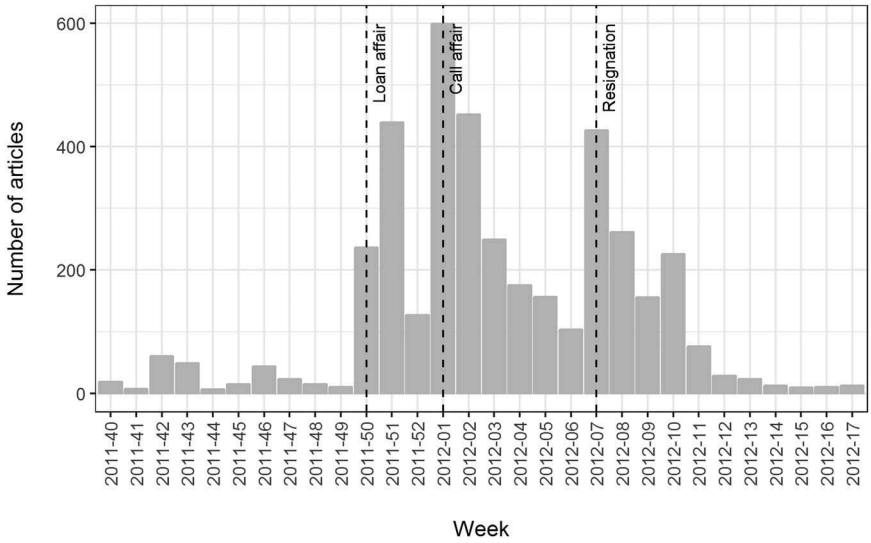

Figure C1. Full distribution of sentiment scores across human codes and manifesto categories.

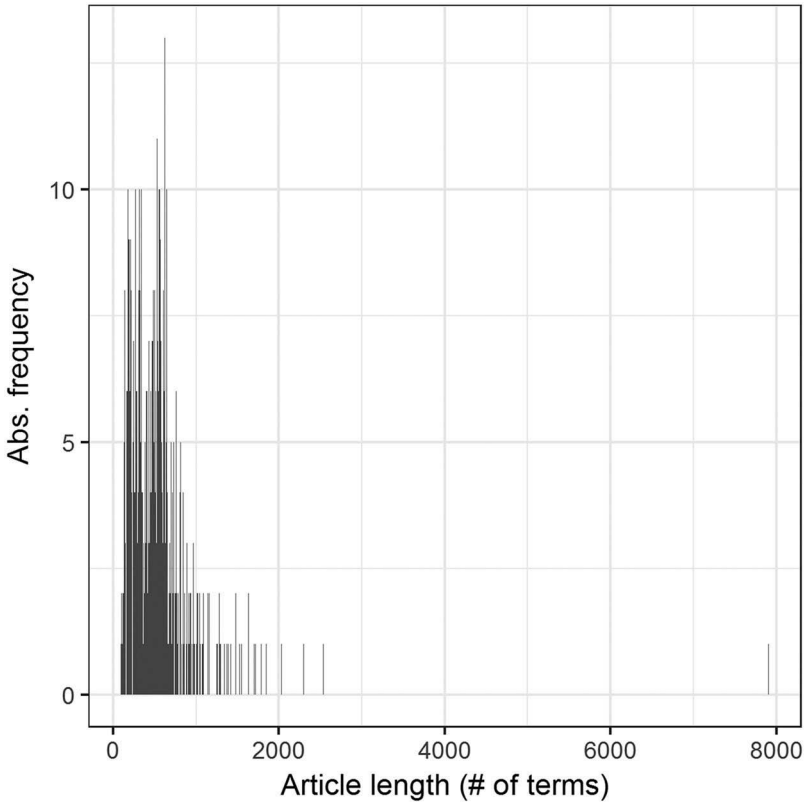

Figure C2. Temporal distribution of newspaper articles in 'Causa Wulff' sample.

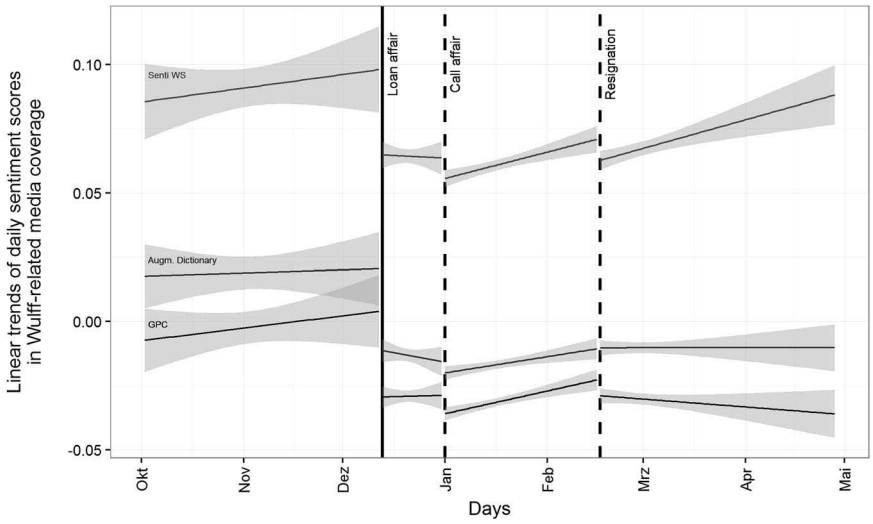

Figure C3. Histogram of article lengths in 'Causa Wullf' sample.