Creating and Comparing Dictionary, Word Embedding, and Transformer-Based Models to Measure Discrete Emotions in German Political Text

Tobias Widmann[1] and Maximilian Wich[2]

[1] *Department of Political Science, Aarhus University, Bartholins Allé 7, 8000 Aarhus, Denmark. Email: widmann@ps.au.dk*
[2] *Department of Informatics, Technical University Munich, Boltzmannstraße 3, 85748 Garching, Germany.*
Email: maximilian.wich@tum.de

Abstract

Previous research on emotional language relied heavily on off-the-shelf sentiment dictionaries that focus on negative and positive tone. These dictionaries are often tailored to nonpolitical domains and use bag-of-words approaches which come with a series of disadvantages. This paper creates, validates, and compares the performance of (1) a novel emotional dictionary specifically for political text, (2) locally trained word embedding models combined with simple neural network classifiers, and (3) transformer-based models which overcome limitations of the dictionary approach. All tools can measure emotional appeals associated with eight discrete emotions. The different approaches are validated on different sets of crowd-coded sentences. Encouragingly, the results highlight the strengths of novel transformer-based models, which come with easily available pretrained language models. Furthermore, all customized approaches outperform widely used off-the-shelf dictionaries in measuring emotional language in German political discourse.

Keywords: text-as-data, emotions, political text, dictionary, word embeddings, transformer models

1. Introduction

Over the last decades, emotions and affect have increasingly moved onto the center stage in political science. Even though citizens have been traditionally regarded as rational actors (Downs 1957) and democratic theory perceived emotions as hindrances to finding optimal solutions (Berelson 1952), recent research emphasizes the inevitability of emotions in political thinking and behavior. Emotional responses influence not only how citizens form beliefs and attitudes, but also whether they participate in politics and who they vote for (Brader 2006; Healy, Malhotra, and Mo 2010; Marcus, Neuman, and MacKuen 2000; Valentino *et al.* 2011; Vasilopoulos *et al.* 2018).

With the rise of the Internet, the availability of political text significantly changed and the analysis of large text datasets is becoming the new normal (Grimmer and Stewart 2013; Soroka, Young, and Balmas 2015). One method widely used by researchers to measure emotional language in text is sentiment analysis (measuring positive and/or negative *valence*). However, language can also engender different types of emotions (Pennebaker and Francis 1996; Roseman, Abelson, and Ewing 1986). Furthermore, research has shown that these different emotions differ starkly in their behavioral effects (e.g., Druckman and McDermott 2008; Lerner and Keltner 2000; Valentino *et al.* 2011; Vasilopoulos *et al.* 2018). This emphasizes the need of moving beyond mere valence toward analyzing language associated with specific emotions. Yet, the availability of emotional dictionaries is highly limited.

Furthermore, until recently, sentiment analysis in social sciences almost exclusively relied on a bag-of-words or dictionary approach. Even though this approach has a number of distinct advantages (fast, cheap, and easy to replicate), it also comes with a series of disadvantages.

***Political Analysis* (2023)**
vol. 31: 626–641
DOI: 10.1017/pan.2022.15

Published
29 June 2022

Corresponding author
Tobias Widmann

Edited by
Jeff Gill

First, dictionaries are often tailored to a specific domain (e.g., e-commerce) and language context (predominantly English). Thus, applying off-the-shelf dictionaries to other domains can lead to poor results (González-Bailón and Paltoglou 2015). Second, bag-of-words approaches analyze words without contextual information. Yet, excluding context leads to a loss of information that otherwise could improve accuracy (Grimmer and Stewart 2013). Novel models relying on word (or sentence) embeddings can overcome this limitation by learning the meaning of terms through co-occurring words. Recent studies applying these approaches show promising (Rheault and Cochrane 2019; Rudkowsky *et al.* 2018), yet the potential of this approach in the field of discrete emotions needs further investigation.

The goal of this study is therefore twofold. First, we set out to create and validate a novel emotional dictionary ("ed8") that moves beyond valence to measure language associated with eight different discrete emotions. Furthermore, this dictionary is specifically tailored to political language in a non-English-language context (German). In a second step, we move beyond the bag-of-words approach and create new tools to measure emotional appeals in political communication: locally trained word embeddings combined with a simple neural network classifier and a transformer-based model (ELECTRA, "Efficiently Learning an Encoder that Classifies Token Replacements Accurately"). To do so, we use approximately 10,000 crowd-coded sentences in German to provide training and test data for the machine learning classifiers. We subsequently compare the performance of the three new tools (ed8 dictionary, word embedding-based neural network classifiers, and transformer-based model) created in this study to freely available off-the-shelf dictionaries. To increase the validity of our tools, we further conduct a series of robustness tests including an additional dataset of crowd-coded sentences and a case study for hypotheses testing.[1]

In doing so, this paper entails a series of important contributions: First, it provides three new tools to measure discrete emotional appeals in political communication. Furthermore, rigorous validation tests show that novel transformer-based models are superior to all other approaches in measuring discrete emotional appeals. This finding is reassuring as it shows that pretrained transformer-based models, which can be easily applied to other languages and domains, outperform costlier and more time-consuming tools in the analysis of political text. Lastly, all three customized tools of this study significantly improve the measurement of emotional language compared to widely used off-the-shelf dictionaries. This last finding emphasizes the need for caution when relying on results computed by ready-to-use dictionaries.

2. Previous Work on Affective Language in Political Text

Automated sentiment analysis of textual data is one way to study the emotive language of political communication. Recent studies suggest that political parties use emotive rhetoric in a strategic manner, depending on their policy positions (Kosmidis *et al.* 2019), the state of the economy (Crabtree *et al.* 2020), or the temporal direction of political statements (Müller 2020). To measure the emotional content of political messages, these studies rely predominantly on sentiment dictionaries that measure positive and negative valence of text using predefined lists of vocabulary. Among the most widely used dictionaries is, for instance, the Linguistic Inquiry Word Count (LIWC) dictionary from the field of psychology (Pennebaker, Francis, and Booth 2001). Yet, there exists an abundance of other dictionaries from different fields (Bradley and Lang 1999; Hu and Liu 2004; Nielsen 2011; Stone *et al.* 1962; Young and Soroka 2012). Even though they differ in the discipline they were created for, these lexica have two things in common: First, they mainly measure positive versus negative valence. Second, they use a bag-of-words approach to measure sentiment, ignoring contextual information of words.

[1] Replication code for this article is available in Widmann and Wich (2021) at https://doi.org/10.7910/DVN/C9SAIX.

2.1. Moving Beyond Valence

Research in political psychology revealed that different emotions, even of the same valence, can influence important political processes differently (Brader 2006; Druckman and McDermott 2008; Lerner and Keltner 2000; Valentino *et al.* 2011; Vasilopoulos *et al.* 2018). Moreover, research has provided proof that discrete emotions can be transmitted through text. Encountering emotionally charged words or emotion-specific appraisal patterns in text can trigger discrete emotional responses which then, in turn, can carry emotion-dependent consequences for information processing, political attitudes, and political behavior (see, e.g., Kühne and Schemer 2015; Nabi 2003). Thus, based on this research, this study contends that it is not enough to simply analyze whether parties or politicians use negative or positive tone. Instead, we argue it is necessary to analyze text for discrete emotional rhetoric since it can lead to distinct political consequences.

Yet, despite their importance, only a small number of studies look at discrete emotions (e.g., Back, Küfner, and Egloff 2011; Soroka *et al.* 2015; Tumasjan *et al.* 2010). Existing studies often fall back on available off-the-shelf dictionaries, for example, the LIWC dictionary. The LIWC includes categories for anger, anxiety, and sadness. The NRC dictionary (Mohammad and Turney 2013), another available off-the-shelf dictionary, includes categories for eight different emotions and feelings. Yet, none of these lexicons are tailored to the analysis of political speeches. Political language, however, uses specific vocabulary with specific interpretations (Rheault *et al.* 2016). While prior research shows that these dictionaries can constitute useful tools to analyze different aspects of political language (Jordan *et al.* 2019; Proksch *et al.* 2019), more thorough investigation is necessary to see whether this is also true in the field of discrete emotions. Furthermore, many dictionaries are created primarily for the English-language context, even though some of them provide translated versions. For instance, the NRC dictionary is available in more than 100 languages, relying on Google Translate for automatic translation. The LIWC dictionary, on the other hand, was manually adapted to other languages by paying close attention to language-specific characteristics and vocabulary (for the German version, see Meier *et al.* 2018).

Nevertheless, the application of dictionaries to other language contexts and domains represents a concern that has been brought up numerous times by social scientists while stressing the need for customized tools (González-Bailón and Paltoglou 2015; Grimmer and Stewart 2013; Haselmayer and Jenny 2017; Rheault *et al.* 2016; Soroka *et al.* 2015; Young and Soroka 2012). Thus, this study will create a novel dictionary ("ed8") that is tailored to political communication in German. Using crowd-coded sentences as a benchmark, we will compare its performance with two widely used off-the-shelf dictionaries that are available for German language and measure specific emotions: the LIWC dictionary and the NRC dictionary.

2.2. Moving Beyond a Bag-of-Words

Prior research measuring affective language in political communication relied predominantly on the bag-of-words approach. Yet, bag-of-words representation of text perceives words as independent of their context and ignores the word order of text. Recent developments in natural language processing, such as word embeddings (Mikolov *et al.* 2013), generated new ways to avoid this form of information reduction.

Word embedding models learn the meaning of words by taking the context of words into consideration, and thus they consider the semantic relations between words. To do so, they transform words into numerical vectors that can be represented in a multidimensional space. Within this space, words that carry similar meaning are positioned closer to each other, and words with dissimilar meaning are positioned further apart. As a result, distances between word vectors become informative about the meaning of words.

Recently, an increasing number of studies use such word embeddings for various applications, such as measuring sentiment (Rudkowsky *et al.* 2018), tracking the changing meaning of political

https://doi.org/10.1017/pan.2022.15 Published online by Cambridge University Press

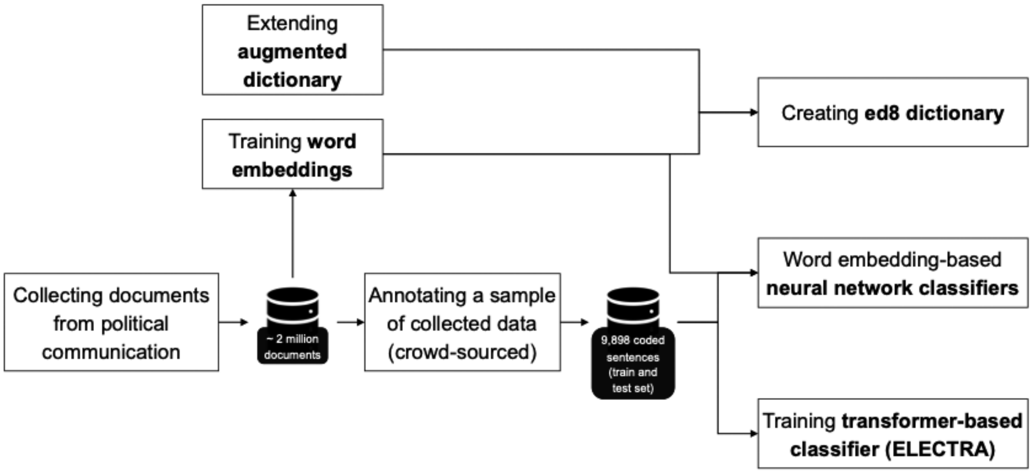
Extending augmented dictionary
Training word embeddings
Creating ed8 dictionary
Word embedding-based neural network classifiers
Collecting documents from political communication
~ 2 million documents
Annotating a sample of collected data (crowd-sourced)
9,898 coded sentences (train and test set)
Training transformer-based classifier (ELECTRA)

Figure 1. Overview over the three different approaches.

concepts over time (Kozlowski, Taddy, and Evans 2019), or measuring partisanship (Rheault and Cochrane 2019). Based on this, the application of word embeddings appears promising, but needs to be investigated and compared more carefully, especially in the field of discrete emotions.

A downside of word embeddings is, however, that a word always has the same representation (embedding) independently from the sentence that it appears in. Let us take the following two sentences as examples: (1) I sit on the bank of the river and (2) I borrow money from the bank. It is obvious that "bank" has two different meanings here. However, the word embedding of "bank" is the same. To address this issue and integrate the actual context of a word, language representation models based on deep neural network architectures were developed. Trained on large amounts of text data, they provide representations of sentences or texts and not only of a word. A prominent representative of these models is bidirectional encoder representations from transformers (BERT), which set new standards when released (Devlin *et al.* 2019).

Since then, transformer-based models are an elementary part of the research in natural language processing, and they provide state-of-the-art performance in several natural language understanding benchmarks, such as paraphrase identification and sentiment analysis (e.g., He *et al.* 2020). In this study, we use both the word embedding and the transformer-based model approach. First, the word embedding approach serves us to extend the ed8 dictionary with new vocabulary. Second, we use the word embeddings and combine them with a simple neural network to train a classification model on crowd-coded training sentences. Third, we train a transformer-based classification model—a more complex neural network—to identify emotional appeals in political communication. Subsequently, we compare all three approaches. Figure 1 illustrates the different approaches in this study.

3. Three Ways to Measure Discrete Emotional Language

3.1. ed8: Creating a Novel Emotional Dictionary

In order to compare the bag-of-words to other approaches, we first need to create a new dictionary tailored to German political communication ("ed8"). The novel ed8 dictionary is capable of measuring language associated with eight different emotions: anger, fear, disgust, sadness, joy, enthusiasm, pride, and hope. Additional information on the importance of these emotions can be found in Online Appendix A.

The ed8 dictionary is based on the "augmented dictionary" (Rauh 2018), a German sentiment dictionary that reliably discriminates between positive and negative tone in German political language. Yet, it cannot differentiate between discrete emotions. Thus, we first extended the

https://doi.org/10.1017/pan.2022.15 Published online by Cambridge University Press

augmented dictionary with emotional categories that attribute words to the eight different emotions mentioned above. All terms have been manually reviewed and—if suitable—attributed to one or more of the different emotional categories. During this step, all terms that carry a clear valence (positive or negative) but are not associated with one of the eight emotions have been dismissed. Important to note is that not all terms are necessarily emotional terms (such as emotionally charged adjectives), but rather words that hint toward the presence of a specific emotional appeal that might be appraised by humans as such. This makes the ed8 dictionary comparably longer than other dictionaries (for the German LIWC, see Meier *et al.* 2018). However, we chose this approach since previous research indicated that discrete emotions cannot only be measured by counting emotional adjectives. Relying predominantly on adjectives has been found to be successful in other classification tasks, yet the classification of discrete emotions requires more situational information (Wang *et al.* 2012).

From the total of 30,070 word forms included in the augmented dictionary, 19,091 terms have been manually assigned to one or more emotional categories in the first step. In a next step, an additional 1,491 emotional terms (including inflections) have been added using word embeddings. Word embeddings represent a convenient way of finding synonyms as the underlying algorithm positions words with similar meaning close to one another in a multidimensional space. Thus, we firstly identified strong emotional words from each category and then used the embeddings to display their 50 "nearest" words, based on their numerical word vectors. From this list of synonyms, we manually selected words that were suitable and not yet part of the dictionary and added these (and their reflections) to the respective category (examples of these words can be found in Online Appendix A).

Thus, the new ed8 dictionary consists of a total of 20,582 terms. Online Appendix A presents information about the length of the individual emotional categories, example terms for each emotion, and more details on negation control. Furthermore, it shows the results of an intercoder reliability test using a trained coder to replicate the attribution of terms to emotional categories on a smaller sample.

Preprocessing steps include the complete removal of numbers and punctuation as well as setting the remaining terms to lower case. Our dictionary does not include word lemmas because we want to build a dictionary that can be used without much effort and independently from computational resources. Integrating more complex Natural Language Processing (NLP) strategies (e.g., lemmatizing, more complex negation rules) requires more complex preprocessing and inferencing steps, going beyond searching and counting occurrences of words and requiring technical skills (Liebeck and Conrad 2015; Wartena 2019).

To calculate the final emotional scores, the word lists are applied to the text corpus to find and count emotional words. To create comparable scores independent of the length of a given document, normalized emotional scores are created, that is, dividing the emotional scores by the word count of each document. We followed the strategy of the augmented dictionary (Rauh 2018) and excluded stop words from the calculation of the normalized emotional scores, rendering the scores more evenly distributed. However, to make sure that the removal of stop words does not bias our results, we replicated parts of the analysis with emotional scores calculated including all terms (see Online Appendix L).

3.2. Creating Word Embeddings and Neural Network Classifiers

3.2.1. *Collecting and Annotating Dataset.* To create word embeddings and train classification models, it is necessary to obtain large text corpora. For this purpose, we collected nearly 2 million German-language documents from political communication of three various countries (Germany, Austria, and Switzerland) and different text sources (Facebook, Twitter, press releases, and parliamentary speeches). The documents have been collected manually, except Bundestag speeches included

https://doi.org/10.1017/pan.2022.15 Published online by Cambridge University Press

in *ParlSpeech* V2 (Rauh and Schwalbach 2020). These types of political communications have been chosen due to their relevance for large parts of the citizenry: Press releases (Schaffner 2006) and parliamentary speeches (Proksch and Slapin 2012) are regularly picked up by news media and reach thereby larger audiences. Furthermore, Facebook and Twitter represent two of the most important social media networks for political discussions. Facebook is by far the largest social network in the German market (Statista 2020). Moreover, Facebook is the main social media platform for political parties in Germany, especially for radical parties (Arzheimer and Berning 2019). Twitter is significantly smaller but often used by political elites to communicate and set the agenda (Barberá *et al.* 2019). The embeddings are therefore trained on a large and diverse dataset of political sentences which should make them more "robust" and less corpus-specific.

This "transformation" dataset is already suited for creating word embedding models after some preprocessing steps of the documents (e.g., lower casing and removing punctuation). In contrast, machine learning classification models—a form of supervised learning—requires coded data. That means that human coders have to annotate the data according to the classification task. The models learn from the annotated data patterns to differentiate between the different classes (e.g., anger and joy).

Our annotated data consist of 10,000 crowd-coded sentences. The 10,000 sentences stem from two important sources of political communication: German parliamentary speeches and German political parties' official Facebook accounts (see Online Appendix B for a detailed description of the data used in the crowd-coding process). The sentences were coded by annotators from a German crowd-working platform called "Crowdguru," which is similar to Amazon's Mechanical Turk. The 10,000 sentences were then compiled into microtasks (human intelligence tasks [HITs]) consisting of 10 sentences each. Every HIT was coded by five different coders, which has been shown to result in enough judgments per sentence to achieve reasonable precision (Benoit *et al.* 2016). Online Appendix B provides the codebook and additional information on quality control mechanisms, the crowd-coding platform, and a discussion on ethical concerns of crowd-sourcing.

The total amount of sentences used in the study is 9,898 sentences, after removing all sentences that have been coded as incomprehensible by two or more coders. Subsequently, these sentences were split in two portions: 90% serve as training data and 10% as test data.

3.2.2. Creating Word Embeddings. To create locally trained word embeddings, we use the "transformation data" described above in order to transform words into their embeddings. This dataset needs to be sufficiently large in order to produce useful results (Spirling and Rodriguez 2022). Consequently, researchers can resort to pretrained embeddings trained on vast amounts of text data. These ready-to-use corpora (e.g., Al-Rfou', Perozzi, and Skiena 2013; Bojanowski *et al.* 2017; Mikolov *et al.* 2017) do not involve any additional computing time and often achieve high accuracy. On the other hand, researchers can also train models locally by using context-specific (e.g., political) data to create the word embeddings. However, this approach can be expensive and time-consuming and therefore not always feasible. Spirling and Rodriguez (2022) compared both approaches and showed that they achieved comparable results. We decided to train word embeddings locally because intuitively the transformation data should be as similar to the corpus of interest as possible. However, we also replicated the analysis with pretrained word embeddings. Online Appendix E presents these additional tests. The findings indicate that advanced pretrained word representations can achieve comparable results as locally trained embeddings.

To locally train the word embeddings, we used the R package *rword2vec*, which implements Google's word2vec algorithm (Mikolov *et al.* 2013). The word2vec algorithm has been widely used in NLP tasks to improve performance of previous approaches (Mikolov *et al.* 2017). The package *rword2vec* embeds each word in 100 dimensions. This means that word distances are computed in a 100-dimensional vector space. Furthermore, we opted for a skip-gram model that

https://doi.org/10.1017/pan.2022.15 Published online by Cambridge University Press

predicts context words given a specific target word. In terms of preprocessing, we transformed the transformation data to lower case and removed links, hashtags, numbers, and punctuation.

After the word embedding models have been trained, we used them in two different ways: to expand the existing ed8 dictionary (described above) and as a way to train simple neural network classifiers.

3.2.3. *Training Simple Neural Network Classifier.* To build our first machine learning model, we followed the procedure of Rudkowsky *et al.* (2018). Before training our actual model, we applied a range of preprocessing steps to convert the test documents into vectors that can be processed by the algorithm. We firstly matched our word embeddings with features in the training dataset. In order to achieve high accuracy, we preprocessed the training data corpus to match as many terms from the training data with the word embeddings. Fewer matching word embeddings per sentence decreases the accuracy of the emotional prediction. Thus, we only removed a small number of words during the preprocessing process (only German stop words), transformed words to lower cases, and used stemming. Afterward, we matched each training sentence with their respective word embeddings. We then averaged all retrieved word embeddings per sentence by calculating the mean vector for each dimension, providing us with sentence embeddings.

After all sentences have been transformed into their corresponding embeddings, a machine learning classifier is applied to learn emotional appeals based on the mean vectors and the human annotation of the training sentences. To do so, we firstly tested a series of different classifiers (Random Forest, Lasso, Naïve Bayes, and Neural Network) which are widely used in statistical learning (James *et al.* 2013). The results of these tests can be found in Online Appendix F. Finally, we opted for a neural network using the *keras* library for R, which achieved the best results. Hyperparameter settings of our neural network models can be found in Online Appendix C. After the neural network models have been trained, we apply the classifiers to the 10% test data and let them evaluate whether a sentence contains emotional appeals or not.

3.3. Training Transformer-Based Classifier

After building classification models with simple neural network architecture, we trained a transformer-based classification model, which is currently state-of-the-art in natural language processing. The transformed-based models are highly complex and very large neural networks. They are pretrained on corpora that contain billions of words, similar to word embeddings. However, in contrast to word embeddings, they contain complex language models that produce contextualized embeddings for entire documents (e.g., one or several sentences). These pretrained models can then be used to train a classification model based on individual data.

We decided to use ELECTRA, an extended BERT version, instead of the classical BERT model (Clark *et al.* 2020). There are two reasons for this decision. First, ELECTRA outperforms comparable BERT models on several benchmarks (Clark *et al.* 2020). Second, the German ELECTRA model that is provided by the German NLP Group and that we use for our study outperforms equivalent BERT models in similar text classification tasks in German.[2] Another difference to the previous architecture is that we train only one model for all emotions and not one model for each emotion. That makes the model easier to use for other researchers.

The model has 12 layers, a hidden state size of 768 and in total 110 million parameters. We used the same train/test split of the data as for the simple neural network architecture. We withheld 10% of the training set as the validation set. In contrast to the previous model, we did not apply any processing steps to the sentences because this is done by the tokenizer of the transformer-based model. We trained the model for four epochs with a learning rate of 5e-5 and chose the best model

[2] The model is provided here: https://huggingface.co/german-nlp-group/electra-base-german-uncased.

https://doi.org/10.1017/pan.2022.15 Published online by Cambridge University Press

of these epochs as the final model. To identify the best model, we defined our own loss function. The problem of the multilabel classification model is that the performances on the different labels can be very unbalanced if one label is easier to be predicted than others. To compensate for this, we defined the loss as follows:

$$\text{Loss} = \sum_{i \in \text{Emotions}} (1 - F1_i) \bullet 2.$$

The double weighting causes that labels that are harder to predict are not neglected. For the transformer-based model, we used the Python library “Transformers” provided by Hugging Face (Wolf *et al.* 2020). Further details on the training process and the hyperparameters can be found in Online Appendix C.

To measure the performance of all three approaches, we calculate precision, recall, and F1 scores. These are typical measurements in machine learning-based classification tasks. Recall is the ratio of correctly predicted observations to the total amount of true observations (indicates the number of false negatives). Precision, on the other hand, is the ratio of correctly predicted observations to the total predicted observations (indicates the number of false positives). The F1 score is defined as the harmonic mean of recall and precision:

$$F1 = 2 \times (\text{Recall} \times \text{Precision}) / (\text{Recall} + \text{Precision}).$$

4. Results

In the first step, we compare the results of the three different approaches while using the human-coded test sentences as “true” data. For an initial comparison, we forced the continuous emotional scores computed by the dictionary to a binary variable, which reflects the output of the two machine learning approaches and the coding decision the human coders faced during the crowd-coding process. A sentence has been perceived as “emotional,” as soon as one human coder coded it as such.

Table 1 shows the results for the three different approaches. The “actual” column shows how many sentences have been judged as “emotional” by human coders (the combined numbers of this column can be greater than the actual size of the test set since one sentence can include multiple emotions). The “predict” column represents the number of sentences classified as “emotional” by the different tools.

As can be seen, there are substantive differences between different emotions and between the three different approaches. Focusing on F1 scores, it becomes obvious that the transformer-based (ELECTRA) model outperforms the dictionary approach and the simple word embedding approach by far. For all emotions under scrutiny, the ELECTRA model shows higher F1 scores compared to the other approaches. The differences are substantively large with the transformer-based model achieving in average 18-point higher F1 scores than the ed8 dictionary. Even though the differences are somewhat smaller, the transformer-based approach still outperforms the locally trained word embeddings approach with F1 scores being in average nine points higher. “Receiving Operating Characteristic” (ROC) curves and confusion matrices for these classifications are reported in Online Appendix G.

The results also indicate that the locally trained word embedding approach outperforms the dictionary approach, even though the differences are not as large as for the ELECTRA model. Furthermore, in Online Appendix E, we replicate this analysis using pretrained word embeddings. As can be seen, the advanced word representations (Bojanowski *et al.* 2017; Mikolov *et al.* 2017) achieve a comparable performance as the locally trained embeddings. This finding shows how easily available pretrained embeddings can achieve better results than tediously created, customized dictionaries.

https://doi.org/10.1017/pan.2022.15 Published online by Cambridge University Press

Table 1. Precision, recall, and F1 scores for the three different approaches.

Emotions	Actual	Predicted	Precision	Recall	F1
ed8 dictionary					
Anger	508	281	0.83	0.46	0.59
Fear	189	287	0.43	0.66	0.52
Disgust	86	182	0.30	0.63	0.40
Sadness	201	289	0.41	0.59	0.48
Joy	143	179	0.46	0.58	0.52
Enthusiasm	220	248	0.44	0.50	0.47
Pride	158	247	0.31	0.48	0.38
Hope	305	303	0.53	0.53	0.53
Word-embeddings-based neural network approach					
Anger	508	500	0.80	0.78	0.79
Fear	189	152	0.61	0.49	0.55
Disgust	86	67	0.60	0.47	0.52
Sadness	201	122	0.70	0.42	0.53
Joy	143	92	0.68	0.44	0.54
Enthusiasm	220	176	0.64	0.51	0.57
Pride	158	123	0.52	0.41	0.46
Hope	305	265	0.69	0.60	0.64
Transformer-based (ELECTRA) approach					
Anger	508	495	0.85	0.83	0.84
Fear	189	221	0.60	0.70	0.64
Disgust	86	89	0.61	0.63	0.62
Sadness	201	181	0.64	0.57	0.60
Joy	143	122	0.70	0.59	0.64
Enthusiasm	220	242	0.62	0.68	0.65
Pride	158	151	0.61	0.58	0.60
Hope	305	352	0.68	0.78	0.73

Looking at differences between emotions, one can observe that some emotions show clearly higher F1 scores compared to others. Anger and hope, for instance, show the highest F1 scores among all emotions for each of the three approaches. These differences can be potentially explained by the higher level of occurrences of these emotions in the training and test data. Online Appendix D shows that anger and hope also exhibit the highest occurrences in the training and test data, as judged by human coders. Figure 2 graphically displays the relationship between the number of occurrences of different emotions and the F1 score of the ELECTRA model.

We test this relationship further by comparing the performance of the different approaches for sentences from different text sources (Facebook posts vs. legislative speeches). We do so because there is reason to expect differences in emotionality between different text types. The results of the comparison between sources are reported in Online Appendix H. Supplementary Table H.1 shows that all three tools exhibit higher performance for Facebook sentences, in comparison to legislative speeches (with the exception of anger for the ELECTRA model). Looking at the emotional occurrences in the test data by text source ("actual" columns in Supplementary Table H.2), it becomes clear that the occurrences are also higher for Facebook sentences (again, with the exception of anger where the numbers are relatively equal). Thus, this finding emphasizes the need for high-quality training and test data for emotion classification, which has been stressed by previous literature (Wang *et al.* 2012).

https://doi.org/10.1017/pan.2022.15 Published online by Cambridge University Press

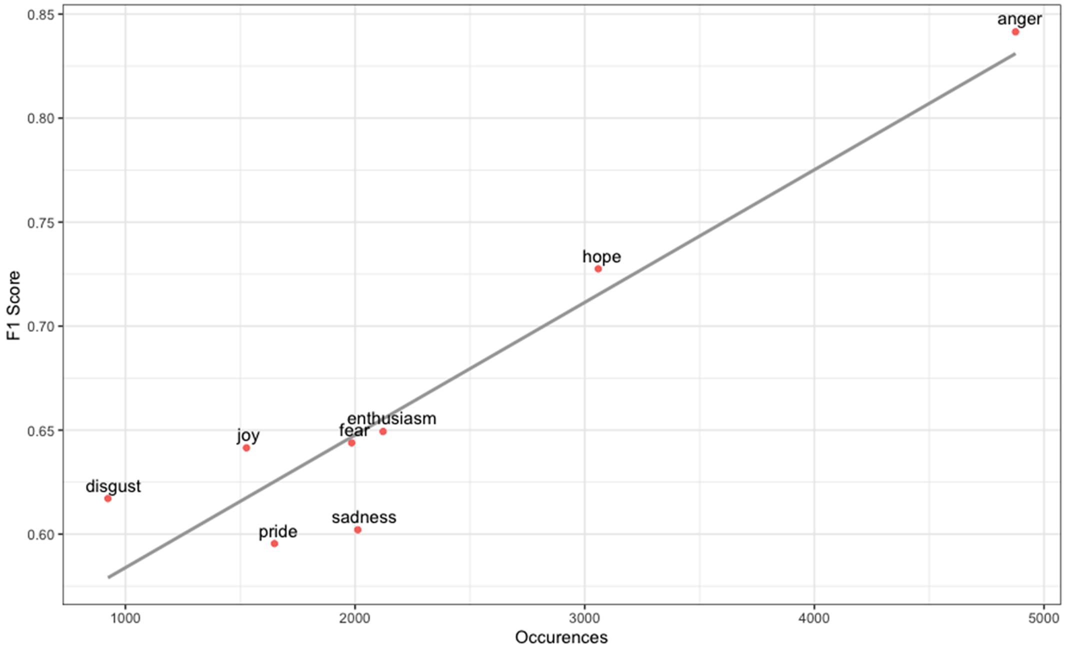

Figure 2. Relationship between level of emotional occurrences and F1 score of the ELECTRA model.

Overall, the main analysis indicates that the transformer-based ELECTRA model achieves by far the best results in measuring discrete emotional appeals. These findings speak for the leveraging of novel deep learning techniques to further improve the accuracy of currently widely used methods in text analysis.

4.1. Off-the-Shelf Dictionaries

In a next step, we compare the results of the newly created tools to freely available dictionaries. As mentioned above, the LIWC and the NRC EmoLex are off-the-shelf dictionaries often applied in political science research. They include not only general categories for positive and negative tone, but also categories for discrete emotions.

We applied both off-the-shelf dictionaries to the 10% test data that we also used to validate the three new tools (including sentences from both Facebook and legislative speeches). The precision, recall, and F1 scores for the LIWC and the NRC dictionaries are shown in Table 2. As can be seen, both dictionaries, LIWC and NRC, show substantively lower F1 scores compared to the novel approaches created in this study. The highest F1 score for the LIWC dictionary is 0.40, for the NRC dictionary 0.25. The automatically translated German version of the NRC EmoLex dictionary shows the lowest scores, with an F1 score of 0.09 for disgust. In Online Appendix I, we present the confusion matrices of this classification. Furthermore, we replicate the same exercise with the complete 9,898 sentences, because a split between training and test data is not necessary for dictionaries. The results remain very similar.

Lastly, we conduct an additional exercise, in which we make use of the continuous scale of the ed8 dictionary in order to see whether higher emotional dictionary scores correlate with stronger agreement on emotions by human coders. The results are reported in Online Appendix I. The graphs illustrate how the ed8 dictionary can significantly discriminate between different categories of human agreement/disagreement.

4.2. Robustness Tests

To test the robustness of the novel tools created in this study, we run a series of robustness tests. First, we replicated the main analysis with a new set of approximately 10,000

https://doi.org/10.1017/pan.2022.15 Published online by Cambridge University Press

Table 2. Precision, recall, and F1 scores for the Linguistic Inquiry Word Count (LIWC) and NRC dictionaries.

Emotions	Actual	Predicted	Precision	Recall	F1
		LIWC dictionary			
Anger	508	178	0.77	0.27	0.40
Fear	189	84	0.40	0.18	0.25
Sadness	201	93	0.48	0.22	0.31
		NRC dictionary			
Anger	508	73	0.81	0.12	0.20
Fear	189	97	0.37	0.19	0.25
Disgust	86	48	0.13	0.07	0.09
Sadness	201	116	0.32	0.18	0.23
Joy	143	64	0.31	0.14	0.19

crowd-codedsentences. This second set of crowd-coded sentences has not been presampled. Instead, it consists of randomly selected sentences from German political Facebook posts and legislative speeches. The reason for this exercise is that presampling using the ed8 dictionary (as in the case of the main analysis) can introduce a strong bias for this specific tool while disadvantaging the off-the-shelf dictionaries. The results of this exercise and additional information can be found in Online Appendix J. The tables indicate that the novel ed8 dictionary is still outperforming freely available off-the-shelf dictionaries. Yet, it becomes obvious that the performance for all dictionaries substantively decreased, which indicates that the results of the main analysis were influenced by the presampling strategy. On the other hand, the superiority of the transformer-based model compared to the remaining approaches becomes even more striking. These results are important as they illustrate the performance of the different tools in a "real-life setting," that researchers applying these tools to new data would encounter.

As a second robustness test, we manipulate the number of coders per crowd-coded sentence. Based on the study by Benoit and co-authors (Benoit *et al.* 2016), we based our main analysis on five crowd judgements per sentence in the first 10,000 sentences. Now, we want to find out whether five judgements per sentence is enough to establish reliable and valid estimates. For the second 10,000 crowd-coded sentences, we therefore took one half (5,000 sentences) aside and let these sentences be coded by 10 crowd coders each. Then we draw on random subsamples from these 5,000 sentences to estimate F1 scores as a function of crowd coders per sentence. We do so by bootstrapping 1,000 sets of subsamples with replacement for each n ranging from of $n = 1$ to $n = 10$ coders per sentence. Then we calculate mean F1 scores for each n and for each approach (ed8, word embeddings, and ELECTRA). Online Appendix K presents the results of this exercise in detail. While, for most emotions, increasing the number of crowd judgments will still slightly improve the F1 scores, we conclude that five crowd judgements per sentence represents a sufficient number, especially when judged from a cost–benefit perspective.

Finally, as a last exercise, we provide an application example which shows how the tools provided in this study can be used for hypothesis testing. In the case study, we analyze more than 12,000 press releases of six German political parties. The results of this exercise are reported in Online Appendix M.

5. Conclusion

This article presents tools to measure discrete emotional appeals in political text. Increased interest in the affective side of politics has led scholars in political science and in political communication to investigate how different political actors use emotional rhetoric in their communication. Yet, a majority of freely available tools measuring emotive language are tailored toward the

https://doi.org/10.1017/pan.2022.15 Published online by Cambridge University Press

English-language context, focus predominantly on positive versus negative sentiment, and rely on the bag-of-words approach.

The approaches presented and validated in this paper move beyond valence to measure discrete emotional appeals. In total, we created and compared three different tools: a novel emotional dictionary (ed8), simple neural network classifiers based on word embeddings, and a transformer-based model. All tools can measure emotional appeals associated with eight discrete emotions. Furthermore, the presented tools are tailored to the German language. Thus, this study adds to the availability of validated tools for measuring discrete emotions in the non-English political context.

Another contribution of this study is that it shows how new transformer-based classification models can be used to analyze political texts regarding their discrete emotional appeals. It therefore adds to a strand of literature that investigates the possibilities of applying novel embedding models in political text analysis (Kozlowski *et al.* 2019; Rheault and Cochrane 2019; Rudkowsky *et al.* 2018). Yet, this study introduces new state-of-the-art NLP models which, to the best of our knowledge, have had little application to date in the analysis of political text. The findings indicate promising results: The validation tests show that the novel transformer-based model clearly outperforms all other approaches (including dictionaries and standard word embedding approaches) for each emotion under scrutiny. It further achieves very good results compared to related, recent text analysis studies in political science (using embeddings to measure sentiment; Rudkowsky *et al.* 2018) or other fields (emotion analysis; Demszky *et al.* 2020; Xu *et al.* 2020). Even though the ed8 dictionary might come with a number of advantages, as it is easier and faster to implement and requires less computing power, it also shows significantly lower performance. Researchers should therefore choose their tool depending on their research goal: While the bag-of-words approach can provide a quick overview of emotional language which, however, requires extensive checking, the ELECTRA model achieves higher accuracy at the price of more resource-intensive application.

Furthermore, the results indicate that the novel tools created in this study achieve significantly higher performance in the classification of discrete emotional appeals compared to freely available off-the-shelf dictionaries. This finding stresses the need for caution when relying on ready-to-use dictionaries. While these dictionaries have been found to perform adequately for some classification tasks (for analytical thinking, see Jordan *et al.* 2019; and for sentiment analysis, see Proksch *et al.* 2019), the findings of this study point toward poor results in the field of discrete emotions. The article therefore reiterates previous calls for customized text analysis tools (González-Bailón and Paltoglou 2015; Grimmer and Stewart 2013; Haselmayer and Jenny 2017; Rheault *et al.* 2016; Soroka *et al.* 2015; Young and Soroka 2012).

In this respect, this study presents encouraging results. While creating new task-specific dictionaries is laborious, the other two approaches can be easily applied to other domains and tasks. First, in regard to the “standard” word embedding approach, additional tests in this article show that advanced pretrained word embeddings (e.g., Bojanowski *et al.* 2017; Mikolov *et al.* 2017) achieve results comparable to the locally trained word embeddings. Thus, when relying on the word embedding approach, scholars do not need to invest time and money to collect large text corpora and compute models, but can instead employ cheap and readily available embeddings. Second, the transformer-based approach, which surpassed all other tools, comes with a pretrained language model that can be easily fine-tuned for a classification task. It is therefore easily applicable to other domains.

In addition, word-embedding-based and transformer-based models are available in a multitude of languages (for Spanish, see Canete *et al.* 2020; for English, see Clark *et al.* 2020; for Swedish, see Malmsten, Börjeson, and Haffenden 2020; and for French, see Martin *et al.* 2019). Even though this article deals with the classification of discrete emotional language in German, it can serve

https://doi.org/10.1017/pan.2022.15 Published online by Cambridge University Press

as a framework to create similar tools for other languages which potentially achieve even better performances. Domain-specific compound nouns, conjugated verbs, and declined words—which are common in the German language but not in other languages (e.g., English)—might decrease the performance of the embedding approaches in this study.

However, we would like to point out that all automated tools presented in this study should be used with substantial caution. As the findings show, there are substantive differences in the ability of the automated approaches to detect specific emotions. While, for some emotions, the tools achieve consistently good results, the detection of others is challenging. Relatedly, the results do not only vary between emotions, but also between text types. Even though we expect variation in the level of emotional appeals between different communication channels, it could also be that emotions are expressed differently in different settings. Yet, machine learning classifiers trained on one specific text type might not necessarily be able to capture these differences. This suggests that researchers might arrive at different results and draw different conclusions when relying on data from different communication channels. This cautionary note does not only apply to emotion detection. Researchers using automated text analysis tools in order to investigate any fine-grained concept need extensive validation steps as the performance of automated methods on new datasets cannot be guaranteed (Grimmer and Stewart 2013). These validation steps could entail the replication of findings using a series of text analysis tools (Schoonvelde, Schumacher, and Bakker 2019) or a more qualitative analysis of the results, by looking at smaller samples of text data. Applying the tools blindly to different texts for different purposes can lead to biased or simply wrong results. This becomes even more urgent, of course, once tools are being transferred to different domains.

The main limitation of this study, which scholars who want to conduct similar analyses should be wary of, concerns the size of the training and test data. In order to make use of the full potential of the different machine learning approaches, researchers need to obtain large sets of test and training data. This study relies on a relatively small sample of emotion-relevant training and test sentences, at least for some emotions. Future research could therefore explore the possibility of automatically created training data to overcome the costs of human annotation (Wang *et al.* 2012). Another limitation of our approach is that our training and test data were annotated on sentence level, and our classification models work on the same level. By doing so, we might have missed emotional appeals caused by the context of the complete text (e.g., the entire speech). Future research could address this by moving from sentence level to paragraph level or document level. The last limitation addresses internal and external validity. Even though we have evaluated our models on additional 10,000 sentences that were not part of the original training and test set (see Online Appendix K), external validity remains limited for two reasons. First, the additional data come from the same sources and can therefore potentially entail the same biases. Second, classification performance (precision, recall, and F1 score) is on average lower on the 10,000 additional sentences. A deviation between the classification performance on the original test set and a new dataset (the additional 10,000 sentences) indicates that the model performs worse in a real-life scenario, meaning the generalizability is somewhat limited. Nevertheless, our results on the new dataset are promising and confirm our approach.

These limitations notwithstanding, this article provides new tools for the research community to analyze emotional rhetoric in political text. It further illustrates how political scientists can use new deep learning methods to improve the accuracy of political text analysis.

Acknowledgments

The authors would like to thank Hanspeter Kriesi and Vicente Valentim for their continuous support throughout the creation of this study. Furthermore, the authors would like to thank Timo Seidl, Johannes Rothe, Kenneth Benoit, participants of the Young Scholars in Computational

https://doi.org/10.1017/pan.2022.15 Published online by Cambridge University Press

Social Science Workshop at the University of Zurich 2020, the editorial team of Political Analysis and the four anonymous reviewers for their help and detailed comments on earlier versions of the manuscript.

Funding Statement

This work was supported by the Early Stage Researcher Grant from the European University Institute, Villa Sanfelice, 50014 San Domenico di Fiesole, Italy.

Conflicts of Interest

There is no conflict of interest to disclose.

Data Availability Statement

Replication code for this article is available in Widmann and Wich (2021) at https://doi.org/10.7910/DVN/C9SAIX.

Supplementary Materials

To view supplementary material for this article, please visit http://doi.org/10.1017/pan.2022.15.

References

Al-Rfou', R., B. Perozzi, and S. Skiena. 2013. "Polyglot: Distributed Word Representations for Multilingual NLP." In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, 183–192. https://arxiv.org/abs/1307.1662

Arzheimer, K., and C. C. Berning. 2019. "How the Alternative for Germany (AfD) and Their Voters Veered to the Radical Right, 2013–2017." *Electoral Studies* 60: 102040. https://doi.org/10.1016/j.electstud.2019.04.004

Back, M. D., A. C. P. Küfner, and B. Egloff. 2011. "'Automatic or the People?': Anger on September 11, 2001, and Lessons Learned for the Analysis of Large Digital Data Sets." *Psychological Science* 22 (6): 837–838. https://doi.org/10.1177/0956797611409592

Barberá, P., et al. 2019. "Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public using Social Media Data." *American Political Science Review* 113 (4): 883–901. https://doi.org/10.1017/S0003055419000352

Benoit, K., D. Conway, B. E. Lauderdale, M. Laver, and S. Mikhaylov. 2016. "Crowd-Sourced Text Analysis: Reproducible and Agile Production of Political Data." *American Political Science Review* 110 (2): 278–295. https://doi.org/10.1017/S0003055416000058

Berelson, B. 1952. "Democratic Theory and Public Opinion." *The Public Opinion Quarterly* 16 (3): 313–330.

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov. 2017. "Enriching WordVectors with Subword Information." Transactions of the Association forComputational Linguistics 5 (June): 135–46. https://doi.org/10.1162/tacl_a_00051.

Brader, T. 2006. *Campaigning for Hearts and Minds: How Emotional Appeals in Political Ads Work*. Chicago: University of Chicago Press.

Bradley, M. M., and P. J. Lang. 1999. "Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings." Technical report C-1, The Center for Research in Psychophysiology.

Canete, J., G. Chaperon, R. Fuentes, and J. Pérez. 2020. "Spanish Pre-Trained Bert Model and Evaluation Data." In *PML4DC at ICLR 2020*. https://users.dcc.uchile.cl/~jperez/papers/pml4dc2020.pdf

Clark, K., M.-T. Luong, Q. V. Le, and C. D. Manning. 2020. "ELECTRA: Pre-Training Text Encoders as Discriminators Rather than Generators." Preprint, arXiv:2003.10555 [Cs].

Crabtree, C., M. Golder, T. Gschwend, and I. H. Indriđason. 2020. "It Is Not Only What You Say, It Is Also How You Say It: The Strategic Use of Campaign Sentiment." *The Journal of Politics* 82 (3): 1044–1060. https://doi.org/10.1086/707613

Demszky, D., D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi. 2020. "GoEmotions: A Dataset of Fine-Grained Emotions." Preprint, arXiv:2005.00547 [Cs].

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." Preprint, arXiv:1810.04805 [Cs].

Downs, A. 1957. *An Economic Theory of Democracy*. New York: Harper.

Druckman, J. N., and R. McDermott. 2008. "Emotion and the Framing of Risky Choice." *Political Behavior* 30 (3): 297–321.

González-Bailón, S., and G. Paltoglou. 2015. "Signals of Public Opinion in Online Communication: A Comparison of Methods and Data Sources." *The ANNALS of the American Academy of Political and Social Science* 659 (1): 95–107. https://doi.org/10.1177/0002716215569192

Grimmer, J., and B. M. Stewart. 2013. “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts.” *Political Analysis* 21 (3): 267–297.

Haselmayer, M., and M. Jenny. 2017. “Sentiment Analysis of Political Communication: Combining a Dictionary Approach with Crowdcoding.” *Quality & Quantity* 51 (6): 2623–2646. https://doi.org/10.1007/s11135-016-0412-4

He, P., X. Liu, J. Gao, and W. Chen. 2020. “Deberta: Decoding-Enhanced Bert with Disentangled Attention.” Preprint, arXiv:2006.03654.

Healy, A. J., N. Malhotra, and C. H. Mo. 2010. “Irrelevant Events Affect Voters’ Evaluations of Government Performance.” *Proceedings of the National Academy of Sciences* 107 (29): 12804–12809. https://doi.org/10.1073/pnas.1007420107

Hu, M., and B. Liu. 2004. “Mining and Summarizing Customer Reviews.” In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 168–177. https://dl.acm.org/doi/abs/10.1145/1014052.1014073

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. *An Introduction to Statistical Learning*, Springer Texts in Statistics, 103. New York: Springer. https://doi.org/10.1007/978-1-4614-7138-7

Jordan, K. N., J. Sterling, J. W. Pennebaker, and R. L. Boyd. 2019. “Examining Long-Term Trends in Politics and Culture Through Language of Political Leaders and Cultural Institutions.” *Proceedings of the National Academy of Sciences* 116 (9): 3476–3481. https://doi.org/10.1073/pnas.1811987116

Kosmidis, S., S. B. Hobolt, E. Molloy, and S. Whitefield. 2019. “Party Competition and Emotive Rhetoric.” *Comparative Political Studies* 52 (6): 811–837. https://doi.org/10.1177/0010414018797942

Kozlowski, A. C., M. Taddy, and J. A. Evans. 2019. “The Geometry of Culture: Analyzing Meaning through Word Embeddings.” *American Sociological Review* 84 (5): 905–949. https://doi.org/10.1177/0003122419877135

Kühne, R., and C. Schemer. 2015. “The Emotional Effects of News Frames on Information Processing and Opinion Formation.” *Communication Research* 42 (3): 387–407.

Lerner, J. S., and D. Keltner. 2000. “Beyond Valence: Toward a Model of Emotion-Specific Influences on Judgement and Choice.” *Cognition & Emotion* 14 (4): 473–493.

Liebeck, M., and S. Conrad. 2015. “IWNLP: Inverse Wiktionary for Natural Language Processing.” In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, edited by C. Zong and M. Strube, 414–418. Beijing: Association for Computational Linguistic. https://doi.org/10.3115/v1/P15-2068

Malmsten, M., L. Börjeson, and C. Haffenden. 2020. “Playing with Words at the National Library of Sweden–Making a Swedish BERT.” Preprint, arXiv:2007.01658.

Marcus, G. E., W. R. Neuman, and M. MacKuen. 2000. *Affective Intelligence and Political Judgment*. Chicago: University of Chicago Press.

Martin, L., B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. de la Clergerie, D. Seddah, and B. Sagot. 2019. “Camembert: A Tasty French Language Model.” Preprint, arXiv:1911.03894.

Meier, T., R. L. Boyd, J. W. Pennebaker, M. R. Mehl, M. Martin, M. Wolf, and A. B. Horn. 2018. “‘*LIWC auf Deutsch’: The Development, Psychometrics, and Introduction of DE-*LIWC2015.” Preprint, PsyarXiv. https://doi.org/10.31234/osf.io/uq8zt

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. “Efficient Estimation of Word Representations in Vector Space.” Preprint, arXiv:1301.3781 [Cs].

Mikolov, T., E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin. 2017. “Advances in Pre-Training Distributed Word Representations.” Preprint, arXiv:1712.09405 [Cs].

Mohammad, S. M., and P. D. Turney. 2013. “Crowdsourcing a Word–Emotion AssociationLexicon.” Computational Intelligence 29 (3): 436–65. https://doi.org/10.1111/j.1467-8640.2012.00460.x.

Müller, S. 2020. “The Temporal Focus of Campaign Communication.” *Journal of Politics* 84: 585–590.

Nabi, R. L. 2003. “Exploring the Framing Effects of Emotion: Do Discrete Emotions Differentially Influence Information Accessibility, Information Seeking, and Policy Preference?” *Communication Research* 30 (2): 224–247.

Nielsen, F. Å. 2011. “A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs.” Preprint, arXiv:1103.2903 [Cs].

Pennebaker, J. W., and M. E. Francis. 1996. “Cognitive, Emotional, and Language Processes in Disclosure.” *Cognition and Emotion* 10 (6): 601–626. https://doi.org/10.1080/026999396380079

Pennebaker, J. W., M. E. Francis, and R. J. Booth. 2001. “Linguistic Inquiry and Word Count: LIWC 2001.” *Mahway: Lawrence Erlbaum Associates* 71 (2001): 2001.

Proksch, S.-O., W. Lowe, J. Wäckerle, and S. Soroka. 2019. “Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Legislative Speeches.” *Legislative Studies Quarterly* 44 (1): 97–131. https://doi.org/10.1111/lsq.12218

Proksch, S.-O., and J. B. Slapin. 2012. “Institutional Foundations of Legislative Speech.” *American Journal of Political Science* 56 (3): 520–537. https://doi.org/10.1111/j.1540-5907.2011.00565.x

Rauh, C. 2018. “Validating a Sentiment Dictionary for German Political Language—A Workbench Note.” *Journal of Information Technology & Politics* 15 (4): 319–343. https://doi.org/10.1080/19331681.2018.1485608

Rauh, C., and J. Schwalbach. 2020. “The ParlSpeech V2 Data Set: Full-Text Corpora of 6.3 Million Parliamentary Speeches in the Key Legislative Chambers of Nine Representative Democracies [Data Set].” Harvard Dataverse. https://doi.org/10.7910/DVN/L4OAKN

Rheault, L., K. Beelen, C. Cochrane, and G. Hirst. 2016. “Measuring Emotion in Parliamentary Debates with Automated Textual Analysis.” *PLoS One* 11 (12): e0168843. https://doi.org/10.1371/journal.pone.0168843

Rheault, L., and C. Cochrane. 2019. “Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora.” *Political Analysis* 28: 112–133. https://doi.org/10.1017/pan.2019.26

Roseman, I., R. P. Abelson, and M. F. Ewing. 1986. “Emotion and Political Cognition: Emotional Appeals in Political Communication.” In *Political Cognition*, edited by R. R. Lau and D. O. Sears, 279–294. Hillsdale, NJ: Lawrence Erlbaum Associates.

Rudkowsky, E., M. Haselmayer, M. Wastian, and M. Jenny. 2018. “More than Bags of Words: Sentiment Analysis with Word Embeddings.” *Communication Methods and Measures* 12: 140–157.

Schaffner, B. F. 2006. “Local News Coverage and the Incumbency Advantage in the US House.” *Legislative Studies Quarterly* 31 (4): 491–511.

Schoonvelde, M., G. Schumacher, and B. N. Bakker. 2019. “Friends with Text as Data Benefits: Assessing and Extending the Use of Automated Text Analysis in Political Science and Political Psychology.” *Journal of Social and Political Psychology* 7 (1): 124–143. https://doi.org/10.5964/jspp.v7i1.964

Soroka, S., L. Young, and M. Balmas. 2015. “Bad News or Mad News? Sentiment Scoring of Negativity, Fear, and Anger in News Content.” *The ANNALS of the American Academy of Political and Social Science* 659 (1): 108–121. https://doi.org/10.1177/0002716215569217

Spirling, A., and P. L. Rodriguez. 2022. “Word Embeddings: What Works, What Doesn’t, and How to Tell the Difference for Applied Research.” *The Journal of Politics* 84: 53.

Statista . 2020. Social Media—Marktanteile der Portale in Deutschland 2020. Statista. https://de.statista.com/statistik/daten/studie/559470/umfrage/marktanteile-von-social-media-seiten-in-deutschland/.

Stone, P. J., R. F. Bales, J. Z. Namenwirth, and D. M. Ogilvie. 1962. “The General Inquirer: A Computer System for Content Analysis and Retrieval Based on the Sentence as a Unit of Information.” *Behavioral Science* 7 (4): 484–498. https://doi.org/10.1002/bs.3830070412

Tumasjan, A., T. O. Sprenger, P. G. Sandner, and I. M. Welpe. 2010. “Predicting Elections with Twitter: What 140.” *Characters Reveal about Political Sentiment*, 8.

Valentino, N. A., T. Brader, E. W. Groenendyk, K. Gregorowicz, and V. L. Hutchings. 2011. “Election Night’s Alright for Fighting: The Role of Emotions in Political Participation.” *The Journal of Politics* 73 (1): 156–170.

Vasilopoulos, P., G. E. Marcus, N. A. Valentino, and M. Foucault. 2018. “Fear, Anger, and Voting for the Far Right: Evidence From the November 13, 2015 Paris Terror Attacks.” *Political Psychology* 40 (4): 679–704. https://doi.org/10.1111/pops.12513

Wang, W., L. Chen, K. Thirunarayan, and A. P. Sheth. 2012. “Harnessing Twitter ‘Big Data’ for Automatic Emotion Identification.” In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing*, 587–592. IEEE. https://doi.org/10.1109/SocialCom-PASSAT.2012.119

Wartena, C. 2019. “A Probabilistic Morphology Model for German Lemmatization.” In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, 40–49. https://serwiss.bib.hs-hannover.de/frontdoor/index/index/docId/1527

Widmann, T., and M. Wich. 2021. “Replication Data for: Creating and Comparing Dictionary, Word Embedding, and Transformer-Based Models to Measure Discrete Emotions in German Political Text [Data Set].” Harvard Dataverse. https://doi.org/10.7910/DVN/C9SAIX

Wolf, T., et al. 2020. “Transformers: State-of-the-Art Natural Language Processing.” In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Association for Computational Linguistics. https://aclanthology.org/2020.emnlp-demos.6

Xu, P., Z. Liu, G. I. Winata, Z. Lin, & P. Fung. 2020. “EmoGraph: Capturing Emotion Correlations using Graph Networks.” Preprint, arXiv:2008.09378 [Cs].

Young, L., and S. Soroka. 2012. “Affective News: The Automated Coding of Sentiment in Political Texts.” *Political Communication* 29 (2): 205–231. https://doi.org/10.1080/10584609.2012.671234

https://doi.org/10.1017/pan.2022.15 Published online by Cambridge University Press