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

Tobias Widmann[1]

Maximilian Wich[2]

ONLINE APPENDIX

[1] Aarhus University, Department of Political Science, Bartholins Allé 7, 8000 Aarhus, Denmark. Email: widmann@ps.au.dk

[2] Technical University Munich, Department of Informatics, Boltzmannstr. 3, 85748 Garching, Germany. Email: maximilian.wich@tum.de

Online Appendix A: Information about the ed8 dictionary

A.1 Selection of Emotions

The emotions under scrutiny include five basic emotions (anger, fear, disgust, sadness, and joy) as defined by Ekman (1999). To these, three additional emotions (pride, enthusiasm, and hope) were added given their central role in political discourse (Brader & Marcus, 2013). The measurement of different discrete emotions is crucial given that they have diverging political consequences.

Anger and fear belong to the most research emotions in the political world. In general, anger and fear have been found to significantly impact risk perception – yet, with diverging effect. Anger, on the one hand, facilitates a more risk-seeking and aggressive behavior, whereas fear promotes risk-averse behavior (Lerner & Keltner, 2000, 2001). Fear and anger have also been found to impact framing effects, the former increasing and the latter dampening the effects of frames (Druckman & McDermott, 2008). Both emotions also impact policy preferences differently, for instance in connection to terrorism and confrontational policies (Huddy et al., 2021; Lerner et al., 2003).

Anger has also been found to strengthen support for new populist radical parties in Europe. Rico et al. (2017) show that in Spain, anger over the economic crisis can make individuals develop populist attitudes. In Britain, anger was associated with higher support for leaving the European Union (Vasilopoulou & Wagner, 2017), and in France, anger is associated with the strengthening of authoritarian policy preferences among right-wing individuals (Vasilopoulos, Marcus, & Foucault, 2018). Other studies found that anger is associated with voting for the French far-right party Front National (Vasilopoulos, Marcus, Valentino, et al., 2018).

In the American context, Banks and Valentino (2012) show that anger is associated with symbolic racism and boosts opposition to racially redistributive policies among white conservatives. Fear, on the other hand, increases conspiracy thinking about minorities (Grzesiak-Feldman, 2013) and makes individuals more likely to search for, remember, and agree with threatening pieces of news about immigrants (Gadarian & Albertson, 2014).

The other emotions included, for instance disgust, also carry important political consequences. From an evolutionary perspective, disgust prompts individuals to stay away from impure, toxic, and dangerous substances. However, research indicates that humans not only experience disgust reactions from physical contamination but also from moral breaches (Lazarus, 1991). People experiencing disgust reactions are more likely to judge moral violations as more severe and have more negative views on specific minority groups (Inbar et al., 2009; Schnall et al., 2008). Moral emotions, such as disgust, can also amplify or even elicit moral judgement of 'immoral behavior' and increase support for the removal of 'immoral behavior' (Desteno et al., 2004). Furthermore, disgust has also been shown to play a role for important political beleifes, such as attitudes towards immigration (Aarøe et al., 2017). Lastly, disgust has been linked to group dynamics, the 'dehumanization' of out-groups and political violence (Matsumoto et al., 2015).

Lastly, sadness has been studied less than other negative emotions in the political context. It has been found to have similar consequences as fear. For instance, individuals have been found to process political information more systematically and rely less on initial bias (Small & Lerner, 2008). Thus, similar to fear, sadness responses make individuals more open to new information and persuasion.

Positive emotions, such as enthusiasm and hope, have been found numerous times to increase political participation and turnout (Brader, 2005, 2006; Just et al., 2007; G. E. Marcus & Mackuen, 1993; Valentino et al., 2011). Hope has been found to be especially important for elections as it links the individual's goals for the future with the democratic process (Kinder, 1994). Enthusiasm has also been found to dampen framing effects (Druckman & McDermott, 2008) and increase reliance on habits and previously held attitudes and beliefs (MacKuen et al., 2010; G. E. Marcus et al., 2000; G. E. Marcus & Mackuen, 1993). Joy, on the other hand, can significantly impact voting decisions (via the perception of incumbents) (Healy et al., 2010). Pride, finally, can be used by politicians and political leaders to create positive ingroup identities, thereby making religious or national identities more salient (Salmela & von Scheve, 2018). This can lead to increased solidarity (Turner, 2007), however, also to prejudice and hostility towards outgroups and minorities (De Zavala et al., 2009)

All in all, based on the studies mentioned above, we believe that the emotions included in the different tools in this study are of high relevance for different aspects of politics.

A.1.1 Differences between emotional categories

Differences between distinct emotions are often explained by underlying appraisal patterns. Appraisals are the dimensions by which individuals evaluate events eliciting emotional responses (Lazarus, 1991). While some emotions differ in terms of valence (anger versus joy), others vary in the level of certainty or situational control that is associated with them, for instance enthusiasm versus hope or anger versus fear (Lazarus, 1991; Smith & Ellsworth, 1985; Tiedens & Linton, 2001). Furthermore, emotions can also differ in their temporal perspective, focusing either on the presence/past (joy) or on the future (hope). These differences in appraisal patterns are reflected in the selection of vocabulary included in the different emotional categories of the ed8 dictionary. For instance, while positive emotions all include positive terms, hope includes more future-oriented vocabulary. In turn, hope includes positive terms that are associated with more uncertainty, as compared to joy and enthusiasm. Examples of emotional words can be found below.

A.2 Dictionary Length and Example Terms

The individual lists of each emotional category have comparable sizes among emotions of the same valence, i.e. all negative emotions have approximately similar lengths and all positive emotions have approximately similar lengths. However, comparing negative categories to positive categories, lists for negative emotions are slightly longer.

The list for anger includes 4743 terms, for fear 4022 terms, for disgust 4212 terms, and for sadness 3885 terms. The list for joy contains 2800 terms, for enthusiasm 2246 terms, for pride 3063 terms, and for hope 2525 terms. Thus, there is an overweight of negative words which, however, might be closer to the real balance of positive and negative terms in German political language (Rauh 2018).

Table A. 1 Example terms for each emotional category

Emotion	Example words
Anger	vicious, evil, brutal, humiliate, damaging, hate, derisive, enraged, …
Fear	terror, anxious, uneasy, uncertain, threatening, punishment, gruesome, …
Disgust	disgusting, repulsive, inhuman, rape, filth, germs, sick, betrayal, corrupt, …
Sadness	miserable, unhappy, mournful, mortifying, humiliating, bleak, drabness, dull, dismal, sad, …
Joy	pleasure, enjoyment, laughter, fabulous, marvelous, praise, pleasant, …
Enthusiasm	fierce, agile, motivated, excited, stimulated, spur, dynamic, energetic, …
Pride	powerful, strong, pride, excellent, rich, glorious, outstanding, paramount, …
Hope	promising, encouraging, forward-looking, healing, hopeful, reanimate, compensation, …

One important note about the definition of disgust: Disgust is generally connected to a physiological process that has evolved to avoid and expel contamination. It prompts individuals to stay away from impure and dangerous substances. However, research indicates that humans not only experience disgust reactions from physical contamination, but also from moral breaches (Lazarus, 1991). Disgust reactions in humans reach "beyond the realm of physical impurity to the realm of moral impurity" (Brader & Marcus, 2013). The disgust category therefore not only includes vocabulary related to physical impurity (e.g. sick, germs, or filth) but also to moral impurity (e.g. corrupt or betrayal).

Table A.2 presents examples of words added through word embeddings (words that were not yet included in the ed8 dictionary, but have been identified via their numerical word vectors):

Table A. 2: Example words that have been identified through their numerical word vectors

word	added to category
forlorn	fear
ruined	sadness
emaciated	sadness
unerring	enthusiasm
grasping	enthusiasm
unswerving	enthusiasm
…	…

A.3 Negation Control

The ed8 dictionary adopts the same negation control as included in Rauh's (2018) augmented dictionary. This negation control recognizes a variety of different negation patterns by including bigram negations of each term. The negation patterns are then replaced by a marker that is subsequently not counted in the final emotional scores. The exact negation pattern in German looks like this:

'(nicht|nichts|kein|keine|keinen) TERM'

A.4 Intercoder-reliability test

To replicate the attribution of emotional words to the individual categories, we made use of a trained expert coder who replicated the task on a smaller sample. The human coder is German native speaker and was briefly trained in differentiating between the distinct emotional categories based on the codebook presented in Online Appendix A. Afterwards, the coder coded a training set consisting of a random sample of 200 emotional words. After assessing the results of this first training for each emotional category, the coder was further trained in the distinction between different emotional categories.

Finally, the human coder categorized another random sample of 250 emotional words. The Krippendorff's Alpha scores for this exercise are in satisfying ranges (see Table A.3), comparable to scores found in similar validation processes (Lind et al., 2017).

Table A. 3: Intercoder-Reliability Estimates (Krippendorff's Alpha)

Anger	.78	Joy	.69
Fear	.69	Enthusiasm	.84
Disgust	.69	Pride	.70
Sadness	.57	Hope	.63

Online Appendix B: Crowd-Coding Process

B.1 Data for Crowd-Coding Process

To conduct the validation process and to receive the ‘true’ answer of the crowd, we used a German crowd-working platform called ‘Crowdguru’, which is similar to Amazon’s Mechanical Turk. As data, we selected 10,000 sentences coming from two important sources of political communication: parliamentary speeches and Facebook. Facebook is by far the most important social network in Germany with a market share above 50 percent, compared to Twitter’s 11% (Statista, 2020). Moreover, Facebook is the main social media platform for political parties in Germany, especially for radical parties (Arzheimer & Berning, 2019; Stier et al., 2017). Legislative speeches, on the other hand, have the potential to reach large audiences through mass media. Previous research shows that journalists regularly pick up on them (Tresch, 2009). One of the biggest German newspapers (Frankfurter Allgemeine Zeitung) covers each parliamentary session by at least one newspaper article in average since the 1950s (Proksch & Slapin, 2012). TV programs also regularly pick up on legislative speeches in news shows (Salmond, 2014). We therefore are convinced, that we chose two important and relevant text types as data sources.

Parliamentary debates stem from the 19th legislative period of the German Bundestag, starting in October 2017 up to June 2019. Facebook posts were taken from the official Facebook accounts of all German parties, during the same period of time. The speeches and Facebook posts were subsequently collapsed into sentences, resulting in a total of 333,572 sentences in parliamentary speeches and 34,375 sentences in Facebook posts. From these sentences, 20 percent of the validation data (2000 sentences) were randomly selected. The remaining 80 percent were selected from pre-sampled data sets. Research on emotional language shows that emotional words only rarely occur in communication (Pennebaker et al., 2003). Thus, randomly selecting sentences from the overall sample would most likely lead to a very low amount of emotionality, which would mean that from the 10,000 sentences only a few hundred sentences would include emotional language. Hence, we applied the ed8 dictionary to the overall sample of sentences and subsequently pre-sampled data sets for each emotion, including only sentences that have an emotional score above 0 for the respective emotion. This resulted in eight data sets (one for each emotion) with sentences that have an emotional score that is greater than 0. From each of these eight pre-sampled data sets we subsequently selected 10 percent (1000 sentences) of the validation sample, resulting in 8000 sentences. Taken together with the 2000 sentences that were randomly chosen, this results in the total validation sample of 10,000 sentences.

B.2 Quality Control

The validation process included different tests before and during the coding process to assure a high quality of coding. First, coders had to finish a start quiz consisting of so-called "gold sentences." Gold sentences are sentences that are clearly associated with one specific emotion and therefore have a pre-defined unambiguous answer (see e.g. Benoit et al., 2016). In the start quiz, 80 percent of the answers had to be answered correctly in order to be admitted to the job. Ten crowd-workers did not pass the start quiz and hence were not allowed to work on the tasks.

During the coding process, the quality of the coders was assured using either gold sentences or 'screeners.' Screener sentences contain a precise instruction on how to label the sentence, which are designed to assure that coders carefully read the sentences (Berinsky et al., 2014). Based on the answers to the randomly chosen test question (gold sentence or screener), trust scores were calculated for each individual coder in regular intervals. Coders whose trust score fell below 60 percent, were ejected from the job and their assignments added back to the HIT pool. However, only one crowd-worker fell below the threshold of 60 percent and was subsequently removed from the task. In total, 69 coders coded the 10,000 sentences.

B.3 Codebook

This section describes the coding instructions that each crowd-coder had to read before starting the coding task on Crowdguru. The following instructions have been translated from German to English.

INSTRUCTIONS:

What is this about? On the following pages we present you several text documents from politics and media, which consist of single sentences. Some of these texts, but not all, express certain emotions, such as anger or joy.

Task: Please read the following sentences slowly and carefully and then decide whether, in your opinion, the texts are connected with one or more of the following emotions:

- Anger
- Fear
- Disgust
- Sadness
- Joy

- Enthusiasm
- Proud
- Hope

Important: You can click on more than just one emotion, since a sentence can be associated with several emotions at the same time. If no emotion is connected to the sentence, please select "no emotion."

Explanation and example sentences:

- Anger: Phrases associated with anger often express displeasure with something or someone, or dissatisfaction and disappointment. They may also contain insults or strong criticism towards a person or a group. For example: "Unbelievable: Murderer kills again, politics remains inactive! How much more stupid can this be? Why is the killer still at large?"
- Fear: Sentences with fear express an oppressive, anxious feeling, a feeling of being threatened or insecure. For example: "Terrible terror in Berlin: The country, shaken by violence, murder and terror, simply does not come to rest. How safe are we still in Germany?
- Disgust: This emotion is often associated with 1) impurity, dirt, or disease 2) or it can also indicate disgusting behavior (in a moral sense).
  - Example 1: "A crime that can hardly be surpassed in disgust: The man has bitten the victim bloody and infected him with hepatitis B."
  - Ex.2: "These statements make us sick. These are sentences that show how corrupt and immoral the government has become in the meantime."
- Sadness: This emotion can be expressed for example by great injustice, suffering, failure, or death. For example: "The humanitarian catastrophe in Yemen is heartbreaking. It is an imperative of humanity that suffering, starvation and death finally come to an end.
- Joy: Sentences expressing joy often emphasize joyful, beautiful things that please oneself or someone else. For example: "What a phenomenal result! Incredible! This is good for Germany, good for democracy and good for the whole country!
- Enthusiasm: Sentences with enthusiasm often describe an excitement for something, an exaggerated dynamism and vitality, or an extreme commitment to a cause. For example: "Those who are committed encourage others. Our society and democracy live on commitment. Thanks to all who participate! Keep up the good work!"

- Pride: Sentences expressing pride often emphasize self-confidence and joy over a possession, a characteristic or a performance. For example: "The performance of our industry and economy is outstanding and admirable. We are a rich and strong country that impresses many!
- Hope: Sentences of hope express confidence in the future, emphasize confidence, optimism about what the future will bring. For example: "I wish everyone a happy, successful and healthy new year. May 2018 bring less war, but more confidence and helpfulness.

What should you pay attention to? Please rate only the content of the text. Please remain impartial, your personal experiences with persons, parties or organizations should not influence your evaluation.

Uncodable: A text can be rated as "uncodable" if the text is incomplete or makes no sense. Sentences or paragraphs may also contain incomprehensible characters because they have been processed automatically. If a rating is impossible, we ask you to classify the text as "uncodable." Example: "Ic&// \n\n\n this!%!, aeut!%%"

Special case: sentences with specific coding instruction

Some texts may contain very specific instructions regarding their encoding. In these cases you should ignore all text contents and follow the instructions only. For example: "And the governing parties continue to remain silent. How long should this continue? Please ignore the content of the previous sentences and encode this text as"uncodable".

Thank you very much for your contribution!

B.4 Information about the crowd-working platform

As mentioned before, the crowd-coding was conducted via a German crowd-working platform called 'Crowdguru'. The company was founded in 2008 and offers services such as providing sales data, picture tagging, classification tasks, content moderation, contact address research and content creation, e.g. for product descriptions. According to website information, the company works with a crowd of more than 45,000 individuals. The majority of services is offered in German. However, according to company information, specific tasks are also available in English, yet, relying on a much smaller English-speaking crowd.

The specific task for this project consisted of sentence classification. Crowd workers were firstly presented with the codebook described above. After passing the start quiz, they could choose to classify as many sentences as they want (as long as they did not fall below the quality threshold).

The costs for the first set of sentences included costs for 10,000 sentences each coded by five different coders (50,000 units) plus test questions (5,000 units). The platform also required a 'set-up fee' for creating and providing the digital infrastructure.

The costs for the second set of sentences included the price for tagging 10,000 sentences, of which half was coded by five coders each (25,000 units) and the other half by ten coders each (50,000 units). In addition, the second crowd-coding included 8,333 test sentences. Since the digital infrastructure was already up and running, no set-up fee was required for the second round

Set	Task	Amount	Price per unit	Total price
First set	Emotion classification of sentence	55,000 units	0.06 €	3,300 €
First set	Setup-up of infrastructure	1 unit	150 €	150 €
Sub-total				**3,450 €**
Second set	Emotion classification of sentence	83,333 units	0.06 €	4,999.98 €
Second set	Setup-up of infrastructure	-	-	-
Sub-total				**4,999.98 €**
TOTAL				**8,449.98 €**

B.5 Discussion of ethical issues related to crowd-coding

While the use of crowd-sourcing for data production in social sciences and other related fields is rapidly growing, it also has become the subject of intense debate over the past years. Reports about unfair payment and other difficult working conditions (Newman, 2019) led to growing concern among researchers about the employment of crowd platforms for social science research. The potential concerns are manifold.

First and foremost, it has been repeatedly reported about the low wages paid to crowd workers (Newman, 2019; O'Connor, 2020). According to a study on crowd workers on Amazon Mechanical Turk (MTurk), the average hourly wage is only $1.25 (Ross et al., 2010). Another study on the same platform calculated a median hourly wage of approximately $2 per hour (Hara et al., 2017). The main reason for these low wages is ‘invisible work’, i.e. work-related tasks that workers have to do but for which they are ultimately not paid for (Hara et al., 2017; Newman, 2019). These include time crowd workers have to spend on finding suitable tasks or waiting for pages to load. Furthermore, workers spend time on defective tasks that ultimately cannot be submitted, and which will therefore not be paid. And lastly, workers can waste time on tasks that can be rejected by requesters and for which they will not receive compensation.

The last point is related to the lack of accountability on the side of the requesters. For crowd workers, MTurk and other platforms do not provide sufficient possibilities to object a requester’s decision or even to communicate with requesters (Hara et al., 2017). Requesters can simply deny the payment of workers without disclosing reasons (Hara et al., 2017; Newman, 2019) which increases insecurities on the side of the workers. And these insecurities are particularly consequential for people who conduct crowd work not as a hobby or for fun, but who are financially dependent on these platforms (Ross et al., 2010; Williamson, 2016). Studies found that a substantive proportion of crowd workers in the US are relying on MTurk to make basic ends meet (Fort et al., 2011; Williamson, 2016). However, these workers are often stripped of basic labor rights that many workers in North American and European countries possess, for instance the right to unionize and to collectively bargain (Fort et al., 2011; Hara et al., 2017).

However, there are ethical concerns beyond the issue of fair payment. Crowd work can be psychologically harmful (Shmueli et al., 2021), especially when dealing with violent textual or visual content such as images of killings or description of accident victims (Newman, 2019). Crowd workers can also inadvertently or subconsciously expose sensitive information about themselves, especially during data production tasks, which can breach the anonymity and privacy of workers (Shmueli et al., 2021). And crowd work can further strengthen imbalances between requesters in industrialized countries and workers in developing countries. For instance, a great proportion of MTurk crowd workers are located in developing countries, such as India or Bangladesh, which increases the risk of including vulnerable populations (Shmueli et al., 2021).

We believe that researchers need to keep these ethical issues in mind when deciding to rely on crowd-sourcing platforms. Researchers can, for instance, opt for companies/platforms that commit themselves to comply with minimum salary and fair and safe labor conditions. A number of German crowd-sourcing companies (among which is also *Crowdguru*, the company used for this study) signed a ‘code of conduct’ (Gebert, 2017) which presents guidelines for a “prosperous and fair cooperation between companies, clients and crowdworkers”. This code entails commitment to ten points: conformance with law, clarification on legal issues, fair payment, provision of motivating and good work, respectful interaction, clear tasks and reasonable timing, freedom and flexibility, constructive feedback and open communication, regulated approval process and rework, and data protection and privacy. It furthermore includes an Ombuds Office which is overseen by one of the largest German labour unions (IG Metall), which also oversees the enforcement of the code of conduct. Crowd workers can use this Ombuds Office as a communication channel to file complaints. The Ombuds Office then decides on cases regarding monetary disputes as well as other matters, including for example platform work processes (Faircrowd.com, 2017).

Nevertheless, we are convinced that this can only be a first step in the right direction. A lack of transparency on the side of crowd-sourcing platforms makes it hard to actually check whether companies keep their self-imposed commitments. A helpful step would therefore be a requirement for companies to publish numbers on actual wages paid to workers. On the other hand, departments and universities could create guidelines for the employment of crowd workers, as they did for other research including human subjects (Williamson, 2016). For instance, the approval of the Institutional Review Board could become mandatory which then could address other concerns beyond payment.

Online Appendix C: Hyperparameter Settings

C.1 Settings for the 'simple' neural network

Neural networks are a simple representation of the human brain. In these models, neurons are interconnected to other neurons creating a network. These neurons are located on layers and data moves through them mostly in only one direction. Our final neural network model consists of a first layer, a hidden layer, and an output layer. The model for each emotion will be trained for 25 epochs. Furthermore, we set apart 10% of the 90% training data as validation data. The model sets apart this fraction of the training data, does not train on it, and evaluate the loss and any model metrics on this data at the end of each epoch. We used the binary cross-entropy function for the model and a 2-dimensional output layer with a softmax activation. The softmax activation function will return the probability that a sentence is associated with a specific emotion or not. We use one hidden layer consisting of different numbers of neurons. During hyperparameter tuning, we tested different numbers of neurons on the first layer (64, 128, 256), the number of neurons of the hidden layer (32, 64, 128), and the dropout rates on the different layers (0.2, 0.3, 0.4). Dropout is a technique where randomly selected neurons are ignored during training (Srivastava et al., 2014). This means that their contribution to the activation of downstream neurons is temporally removed. This should help in preventing overfitting and in creating a network that is capable of better generalization. Below we present the best parameter setting for each emotion model.

Anger: 128 neurons on the first layer, 64 neurons on hidden layer; 40% drop-out rate; Adam optimizer
Fear: 256 neurons on the first layer, 128 neurons on hidden layer; 40% drop-out rate; Adam optimizer
Disgust: 128 neurons on the first layer, 64 neurons on hidden layer; 40% drop-out rate; Adam optimizer
Sadness: 128 neurons on the first layer, 64 neurons on hidden layer; 40% drop-out rate; Adam optimizer
Joy: 64 neurons on the first layer, 32 neurons on hidden layer; 40% drop-out rate; Adam optimizer

Enthusiasm: 256 neurons on the first layer, 128 neurons on hidden layer; 40% drop-out rate; Adam optimizer
Pride: 256 neurons on the first layer, 128 neurons on hidden layer; 40% drop-out rate; Adam optimizer
Hope: 64 neurons on the first layer, 32 neurons on hidden layer; 40% drop-out rate; Adam optimizer

C.2 Settings for the Electra model

The basis of our transformer-based model is the pre-trained German ELECTRA Base model provided by the German NLP Group[3]. To fine-tune this model for the emotion classification task, we used the Python library Transformers provided by Huggingface (Wolf et al., 2020). The hyperparameters are the following: we set the initial learning rate to 5e-5 with 250 warmup steps and a weight decay of 0.01. The batch size for training and evaluation was 32. We trained the model for four epochs and select the model with the lowest value of our custom loss function, which we defined in 3.3. Other hyperparameters were the default values of the Trainer class provided by the library.

[3] https://huggingface.co/german-nlp-group/electra-base-german-uncased

Online Appendix D: Descriptives

In Online Appendix D we rely on the randomly sampled dataset as it portrays a more realistic picture of emotionality in political communication and the correlations between different emotions.

D.1: Level of emotionality

Table D. 1: Number of emotional sentences as judged by human coders per emotion

Emotion	Number of occurrences
Anger	4860
Fear	1417
Disgust	314
Sadness	1233
Joy	1204
Enthusiasm	2909
Pride	1673
Hope	3000

Table D. 2: Number of emotional sentences as judged by human coders per emotion by text source

Emotion	Facebook	Parliamentary Speeches
Anger	2464	2396
Fear	720	697
Disgust	182	132
Sadness	600	633
Joy	608	596
Enthusiasm	1487	1422
Pride	849	824
Hope	1508	1492

D.2: Correlations

Table D.3 to D.6 show the correlations between emotions as measured by the different tools and as judged by the human annotators. All in all, we believe there is enough discriminant validity between the different emotions. The machine learning approaches show in general lower correlations, i.e. they can discriminate better between the different emotions. This might be due to the overlap of vocabulary between different emotional categories in the ed8 dictionary. The table shows weak to moderate correlations for most negative emotions. The exception is the correlation between anger and fear. Here, we can see a stronger correlation than between other negative emotions. However, this is not particularly surprising as they are not expected to be orthogonal. Instead, empirical evidence shows that both emotions are highly correlated and often co-occur (G. Marcus et al., 2017; G. E. Marcus et al., 2006). The anger-fear correlation is also comparable to other findings in previous research (Alhuzali & Ananiadou, 2021) and reflects a similar correlation strength as measured in self-reports (Lerner & Keltner, 2001).

In terms of positive emotions, most correlations are relatively weak, in particular between joy and enthusiasm. The only exception is the correlation between enthusiasm and hope. This is also not surprising since both emotions are forward looking, prospective emotions that often link the presence to the future and thereby provide feedback for current activities (Brader & Marcus, 2013; G. E. Marcus et al., 2000; G. E. Marcus & Mackuen, 1993). Prior research also often combined these emotions as they often co-occur (see e.g. Valentino et al., 2011).

Table D. 3: Pearson correlation matrix of normalized emotional scores (as measured by the ed8 dictionary)

	Anger	Fear	Disgust	Sadness	Joy	Enthusiasm	Pride	Hope
Anger	1,000							
Fear	0,504	1,000						
Disgust	0,351	0,342	1,000					
Sadness	0,394	0,487	0,326	1,000				
Joy	-0,027	-0,047	-0,008	-0,031	1,000			
Enthusiasm	0,021	0,040	-0,032	-0,027	0,046	1,000		
Pride	-0,022	-0,029	-0,030	-0,019	0,373	0,227	1,000	
Hope	-0,044	-0,023	-0,026	-0,014	0,128	0,440	0,300	1,000

Table D. 4: Pearson correlation matrix of normalized emotional scores (as measured by the word embedding approach)

	Anger	Fear	Disgust	Sadness	Joy	Enthusiasm	Pride	Hope
Anger	1,000							
Fear	0,230	1,000						
Disgust	0,147	0,227	1,000					
Sadness	0,185	0,314	0,279	1,000				
Joy	-0,179	-0,071	-0,038	-0,060	1,000			
Enthusiasm	-0,222	-0,092	-0,055	-0,086	0,015	1,000		
Pride	-0,212	-0,092	-0,051	-0,081	0,195	0,147	1,000	
Hope	-0,264	-0,133	-0,084	-0,121	0,020	0,418	0,145	1,000

Table D. 5: Pearson correlation matrix of normalized emotional scores (as measured by the transformer-based model)

	Anger	Fear	Disgust	Sadness	Joy	Enthusiasm	Pride	Hope
Anger	1,000							
Fear	0,350	1,000						
Disgust	0,193	0,277	1,000					
Sadness	0,264	0,435	0,388	1,000				
Joy	-0,279	-0,120	-0,058	-0,113	1,000			
Enthusiasm	-0,336	-0,156	-0,086	-0,161	-0,080	1,000		
Pride	-0,342	-0,148	-0,072	-0,132	0,501	0,201	1,000	
Hope	-0,392	-0,202	-0,130	-0,242	-0,050	0,618	0,073	1,000

Table D. 6: Pearson correlation matrix of normalized emotional scores (as judged by human annotators)

	Anger	Fear	Disgust	Sadness	Joy	Enthusiasm	Pride	Hope
Anger	1,000							
Fear	0,264	1,000						
Disgust	0,163	0,170	1,000					
Sadness	0,256	0,210	0,186	1,000				
Joy	-0,263	-0,113	-0,062	-0,111	1,000			
Enthusiasm	-0,296	-0,141	-0,098	-0,182	0,122	1,000		
Pride	-0,325	-0,139	-0,076	-0,138	0,437	0,214	1,000	
Hope	-0,318	-0,135	-0,099	-0,171	0,125	0,414	0,163	1,000

Online Appendix E: Replicating the main analysis with pre-trained word embeddings

In order to obtain high-quality word-vector representations that can be successfully used for text analysis tasks, researchers are dependent on large text corpora. This, however, makes it a time-consuming and cumbersome task to locally train word embeddings because one needs to collect large text datasets and invest time to compute models. Therefore, it is a convenient solution to rely on pre-trained word representations which are estimated from large text corpora such as news collections, Wikipedia or web crawl.

One of the first freely available collection of pre-trained word embeddings was 'Polyglot' (Al-Rfou' et al., 2013), which offered distributed word representations for more than 100 languages trained on corresponding datasets of Wikipedia articles. However, compared to more recent embeddings, these early-stage word embeddings included word representations with fewer dimensions. Polyglot represents each word in 64 dimensions. Mikolov and colleagues' (2017) word embeddings trained on Wikipedia articles and web crawl include words represented in 300 dimensions. Similarly, 'wiki word vectors' by Bojanowksi and colleagues (2017) represent each word in 300 dimensions as well. Furthermore, both of these word representations improved early embeddings by relying on a number of improvements, such as including subword information (Bojanowski et al., 2017; Mikolov et al., 2017), position dependent features (Mikolov et al., 2017; Mnih & Kavukcuoglu, 2013) and using phrase representations (Bojanowski et al., 2017; Mikolov et al., 2013).

In the following exercise, we replicated the main analysis, which relies on locally trained word embeddings using German political communication, by combining the above described pre-trained word embeddings with the neural network classifier. Then we retrained these classifiers using the same training and test datasets as in the main analysis. The results of this exercise can be seen below in Table E1 to E3.

The results indicate that the locally trained word embeddings clearly outperform the pre-trained word representations from the 'Polyglot' collection (Al-Rfou' et al., 2013). Each of the individual F1 scores for the locally trained word embeddings are higher compared to the Polyglot embeddings, with large differences for certain emotions (e.g. fear, disgust, enthusiasm). This finding clearly speaks for the superiority of locally trained representations.

However, the comparison of the locally trained embeddings to the more recently released pre-trained embeddings is less clear-cut. The Bojanowski et al. embeddings achieve higher F1 scores for fear, disgust, and sadness compared to the locally trained embeddings and lower ones for anger, joy, enthusiasm, pride, and hope. However, the differences in F1 scores for the different emotions are relatively small, with for instance only 1-point difference for anger and sadness. Similar findings can be seen for the comparison between the Mikolov et al. embeddings and the locally trained word representations. The F1 scores for anger are equal, and the Mikolov et al. embeddings outperform the locally trained ones in disgust and sadness. For the other emotions, the locally trained embeddings achieve better performance, yet again, differences are not as striking as for the polyglot embeddings. This finding is further illustrated in the ROC curves in Online Appendix G, which illustrate the comparable performance of recent pre-trained word embeddings (Bojanowski et al., 2017; Mikolov et al., 2017) compared to locally trained embeddings.

Table E. 1: Replicating the main analysis with pre-trained word embeddings combined with a neural network classifier (Polyglot)

Emotions	Actual	Predicted	Precision	Recall	F1
Anger	508	520	0.69	0.71	0.70
Fear	189	127	0.47	0.32	0.38
Disgust	86	50	0.50	0.29	0.37
Sadness	201	114	0.57	0.32	0.41
Joy	143	73	0.59	0.30	0.40
Enthusiasm	220	114	0.57	0.30	0.39
Pride	158	85	0.53	0.28	0.37
Hope	305	224	0.62	0.45	0.52

Table E. 2: Replicating the main analysis with pre-trained word embeddings combined with a neural network classifier (Bojanowski et al. 2017)

Emotions	Actual	Predicted	Precision	Recall	F1
Anger	508	504	0.78	0.77	0.78
Fear	189	197	0.57	0.59	0.58
Disgust	86	97	0.55	0.62	0.58
Sadness	201	138	0.66	0.45	0.54
Joy	143	78	0.69	0.38	0.49
Enthusiasm	220	235	0.53	0.57	0.55
Pride	158	90	0.58	0.33	0.42
Hope	305	274	0.65	0.59	0.62

Table E. 3: Replicating the main analysis with pre-trained word embeddings combined with a neural network classifier (Mikolov et al. 2017)

Emotions	Actual	Predicted	Precision	Recall	F1
Anger	508	495	0.80	0.78	0.79
Fear	189	141	0.63	0.47	0.54
Disgust	86	66	0.67	0.51	0.58
Sadness	201	196	0.57	0.56	0.56
Joy	143	104	0.60	0.43	0.50
Enthusiasm	220	146	0.62	0.41	0.49
Pride	158	100	0.57	0.36	0.44
Hope	305	289	0.63	0.60	0.62

Online Appendix F: Replication of the main analysis with different machine learning algorithms

To test the performance of different machine learning algorithms, we replicated the main analysis combining the locally trained word embeddings with different classifiers often used in statistical learning and the analysis of political text (Imai & Khanna, 2016; James et al., 2013; Muchlinski et al., 2016; Stewart & Zhukov, 2009). These algorithms include ‘random forest’, ‘lasso’, and ‘naïve bayes’. As can be seen, the neural network classifier, used in the analysis in the main text, achieves the highest performance. The differences to the neural network classifier can be relatively small, as in the case of the Naïve Bayes algorithm, but sometimes also substantively large as in the case of the lasso classifier, which only achieves an F1 score of 0.09 for the classification of pride, for instance.

Table F. 1: Replicating the main analysis by combining locally trained word embeddings with a Random Forrest classifier

Emotions	Actual	Predicted	Precision	Recall	F1
Anger	508	496	0.79	0.77	0.78
Fear	189	77	0.66	0.27	0.38
Disgust	86	32	0.66	0.24	0.36
Sadness	201	82	0.75	0.30	0.43
Joy	143	57	0.70	0.28	0.40
Enthusiasm	220	92	0.67	0.28	0.40
Pride	158	38	0.55	0.13	0.21
Hope	305	202	0.71	0.47	0.57

Table F. 2: Replicating the main analysis by combining locally trained word embeddings with a Lasso classifier

Emotions	Actual	Predicted	Precision	Recall	F1
Anger	508	515	0.76	0.77	0.76
Fear	189	82	0.61	0.26	0.37
Disgust	86	24	0.58	0.16	0.25
Sadness	201	87	0.64	0.28	0.39
Joy	143	56	0.75	0.29	0.42
Enthusiasm	220	107	0.64	0.31	0.42
Pride	158	15	0.53	0.05	0.09
Hope	305	211	0.64	0.45	0.53

Table F. 3: Replicating the main analysis by combining locally trained word embeddings with a Naïve Bayes classifier

Emotions	Actual	Predicted	Precision	Recall	F1
Anger	508	632	0.66	0.82	0.73
Fear	189	436	0.32	0.74	0.45
Disgust	86	310	0.21	0.74	0.32
Sadness	201	409	0.36	0.74	0.49
Joy	143	164	0.40	0.46	0.43
Enthusiasm	220	336	0.44	0.67	0.53
Pride	158	330	0.29	0.61	0.39
Hope	305	427	0.53	0.74	0.62

Online Appendix G: ROC curves and confusion matrices for the main analysis

G.1 Confusion matrices

Table G. 1: Confusion Matrices for the ed8 Dictionary

		True			True	
	Anger	0	1	**Joy**	0	1
Predicted	0	434	275	0	751	60
	1	48	233	1	96	83
	Fear	0	1	**Enthusiasm**	0	1
Predicted	0	638	65	0	631	111
	1	163	124	1	139	109
	Disgust	0	1	**Pride**	0	1
Predicted	0	776	32	0	661	82
	1	128	54	1	171	76
	Sadness	0	1	**Hope**	0	1
Predicted	0	622	83	0	544	143
	1	167	118	1	141	162

Table G. 2: Confusion Matrices for the Word Embeddings Approach

		True			True	
	Anger	0	1	**Joy**	0	1
Predicted	0	380	110	0	818	80
	1	102	398	1	29	63
	Fear	0	1	**Enthusiasm**	0	1
Predicted	0	742	96	0	706	108
	1	59	93	1	64	112
	Disgust	0	1	**Pride**	0	1
Predicted	0	877	46	0	773	94
	1	27	40	1	59	64
	Sadness	0	1	**Hope**	0	1
Predicted	0	752	116	0	603	122
	1	37	85	1	82	183

Table G. 3: Confusion Matrices for the Transformer-based Approach (Electra)

		True			True	
	Anger	0	1	**Joy**	0	1
Predicted	0	409	86	0	810	58
	1	73	422	1	37	85
	Fear	0	1	**Enthusiasm**	0	1
Predicted	0	712	57	0	678	70
	1	89	132	1	92	150
	Disgust	0	1	**Pride**	0	1
Predicted	0	869	32	0	773	66
	1	35	54	1	59	92
	Sadness	0	1	**Hope**	0	1
Predicted	0	723	86	0	572	66
	1	66	115	1	113	239

G.2 ROC curves

This section presents the ‘receiver operating characteristic’ (ROC) curves, showing how sensitivity (i.e., true positive rate) and specificity (i.e., true negative rate) of the predictions of the different machine learning approaches vary across cutoffs for the predicted probability. Overall, the ROC curve is a summary statistic about how well a binary classifier performs for the classification task. If the classifier predictions would be unrelated with the binary outcome, the expected ROC curve is simply the y=x line. This would be the worst possible performance of a classifier. In a situation where the classifier can perfectly predict the outcome variable, the ROC curve consists of a vertical line (x=0) and a horizontal line (y=1). All prediction models should lie somewhere in between these two extremes. Higher predictive power of a classifier is illustrated by a curve that is shifted more to the ‘north-west’ of the plot.

The following plots are based on predictions for the test data (10 percent) from models fitted to the observations in the training data (90 percent). The graphs illustrate the performance for two of the three novel machine learning approaches: Word embeddings combined with a simple neural network and the transformer-based approach. For comparison, the graphs also include the performance of pre-trained word embeddings (Al-Rfou’ et al., 2013; Bojanowski et al., 2017; Mikolov et al., 2017) and the performance of the locally trained word embeddings combined with a Naïve Bayes classifier instead of a neural network.

Overall, as can be seen, the graphs illustrate the superiority of the state-of-the-art transformer-based Electra models compared to all other approaches. In all plots, the green curve shows the largest “Area Under the Curve” (AUC) which illustrates that these models are the best in distinguishing between 0 and 1 (not emotional and emotional). For most emotions, the transformer-based models are followed by the locally trained word embeddings (black curve). Their performance, however, is in comparable range with advanced pre-trained word embeddings such as the word representations from Bojanowski and colleagues (red curve, 2017) or Mikolov and colleagues (blue curve, 2017). The worst performance is shown by the pink curve, representing the classification based on Polyglot word embeddings (Al-Rfou’ et al., 2013).

Figure G. 1: ROC curve for different classification models (anger)

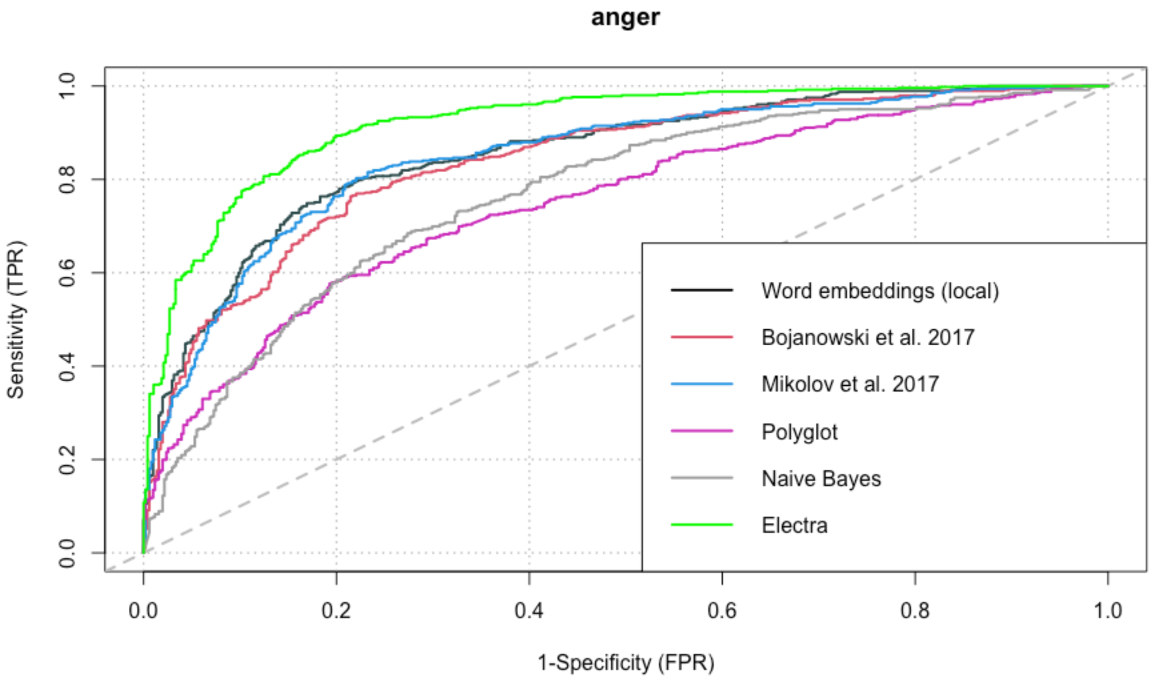


Figure G. 2: ROC curve for different classification models (fear)

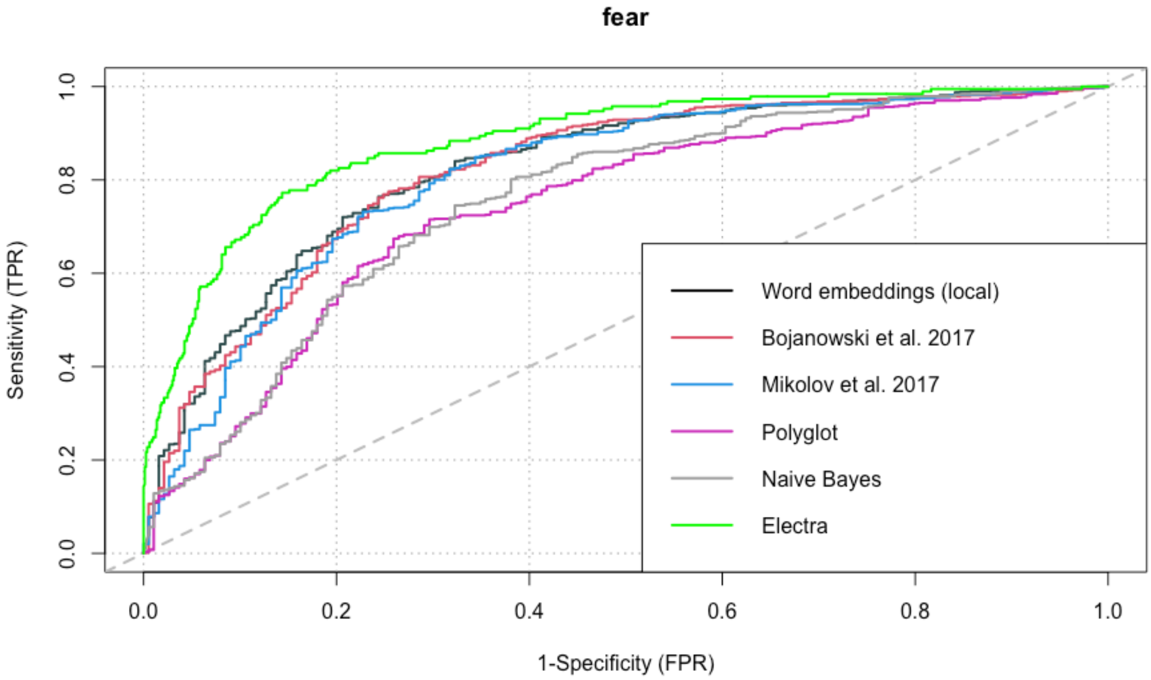


Figure G. 3: ROC curve for different classification models (disgust)

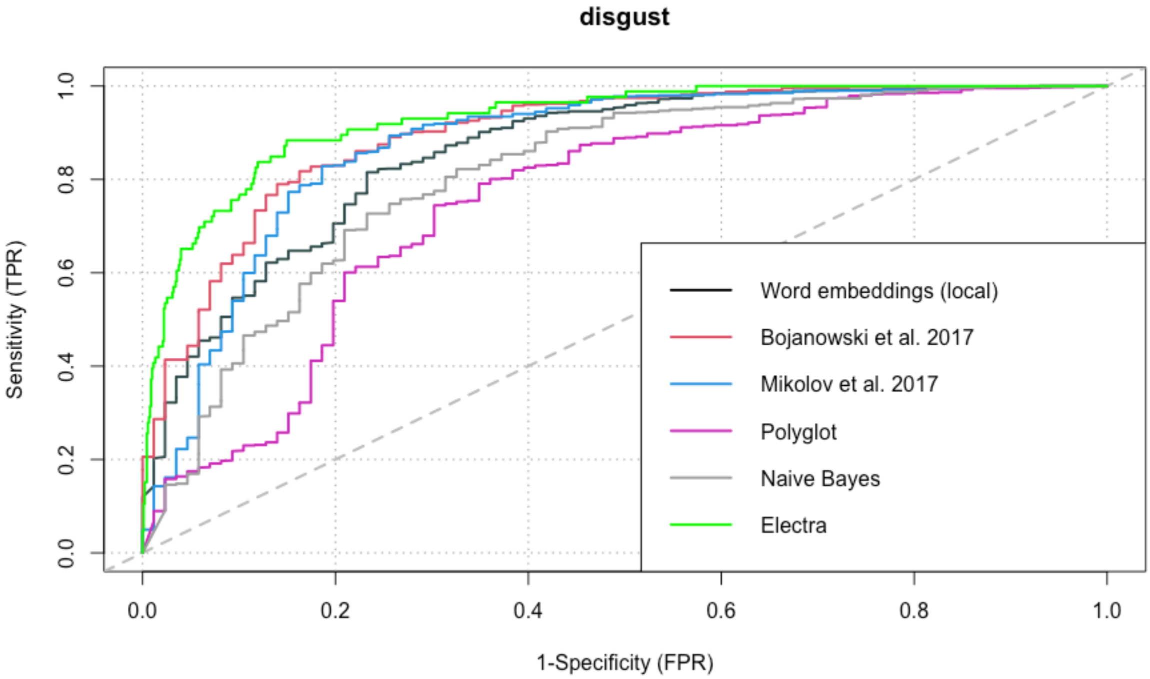


Figure G. 4: ROC curve for different classification models (sadness)

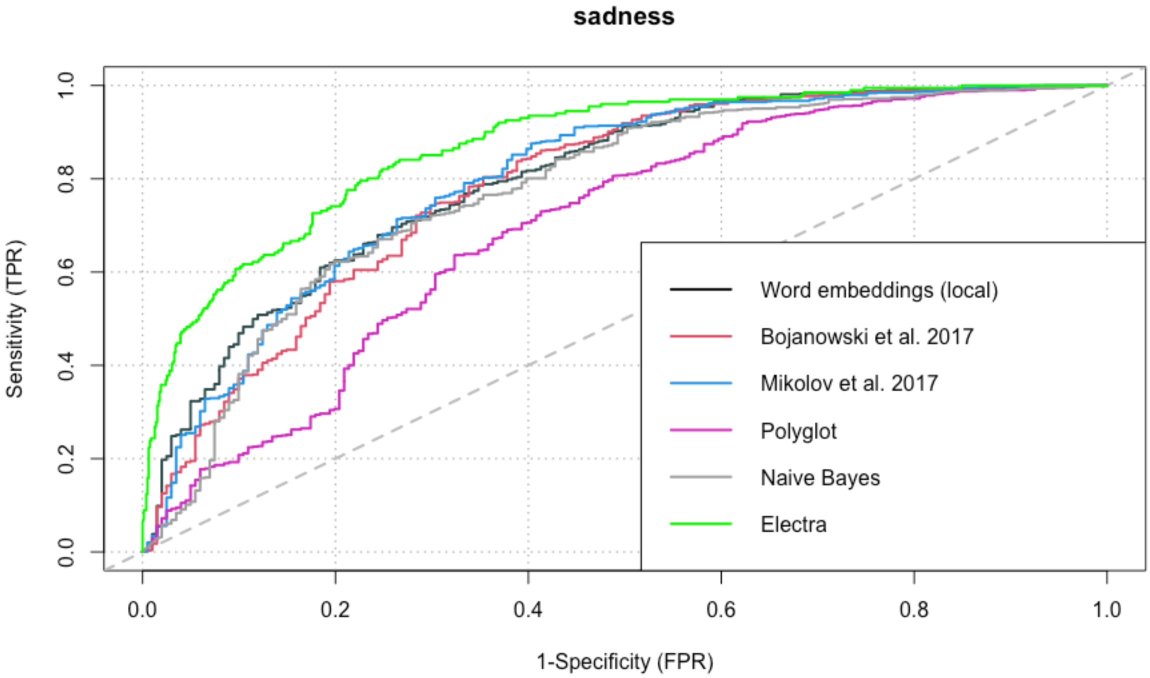


Figure G. 5: ROC curve for different classification models (joy)

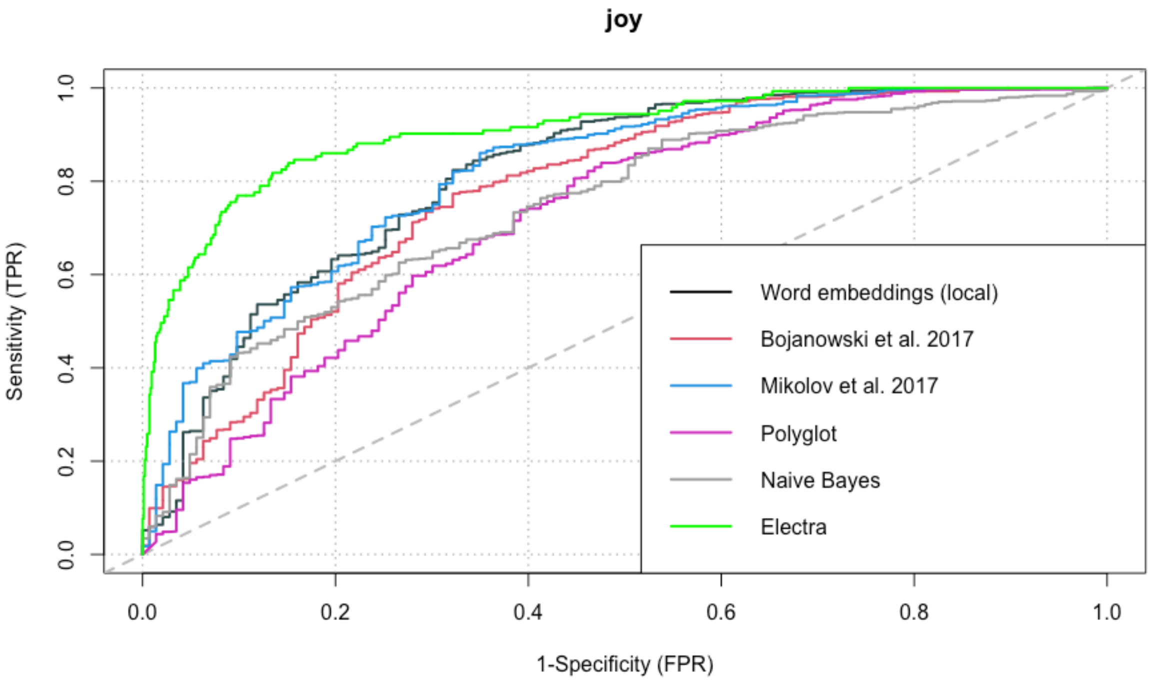


Figure G. 6: ROC curve for different classification models (enthusiasm)

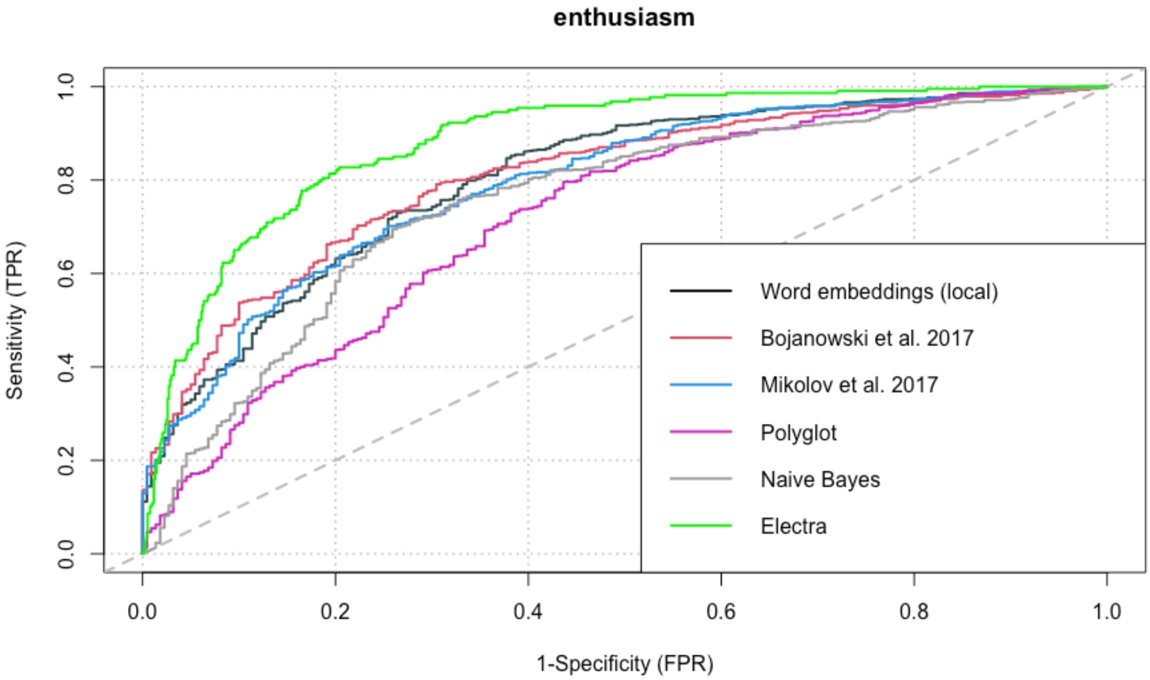


Figure G. 7: ROC curve for different classification models (pride)

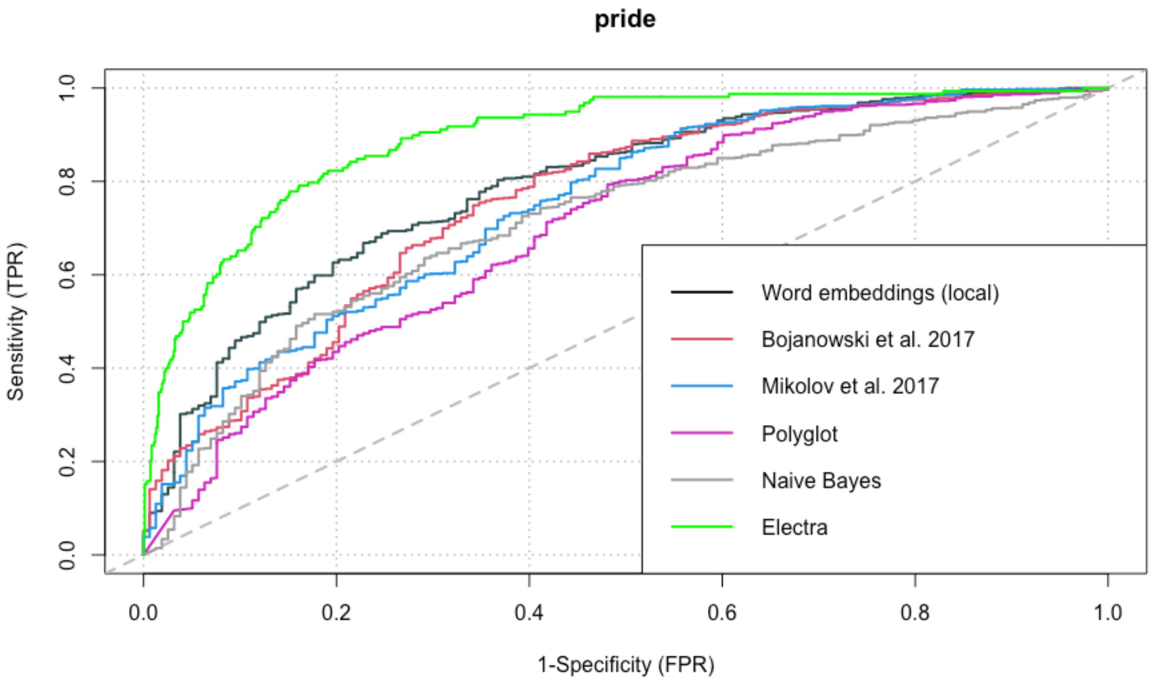
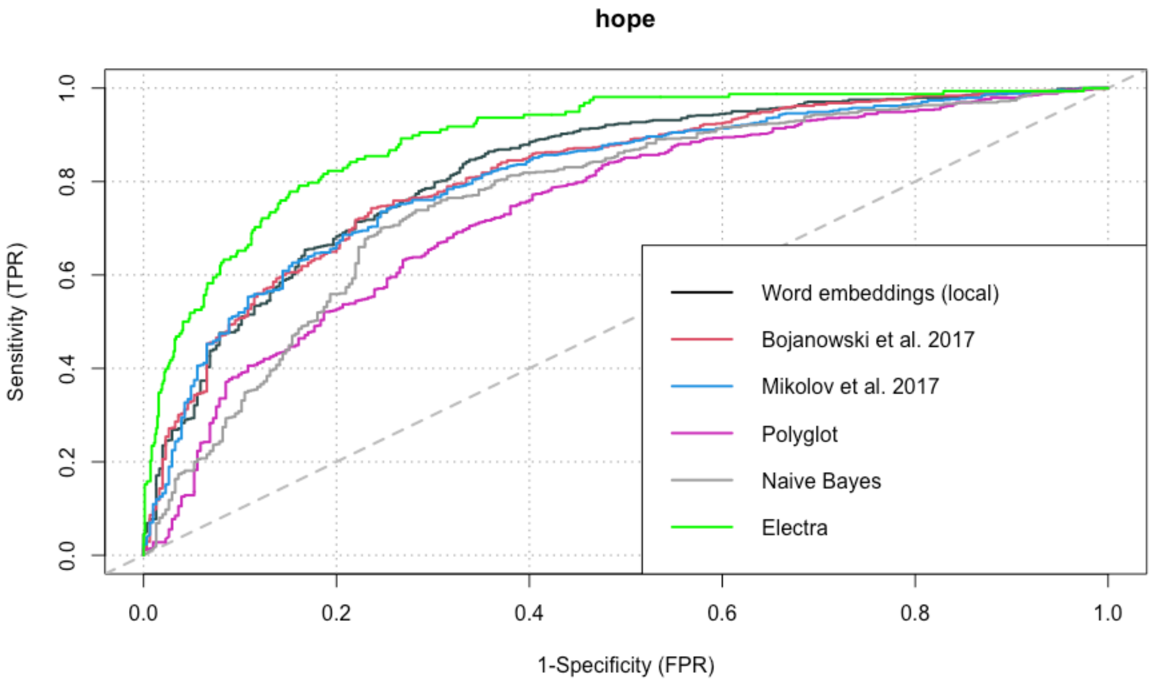


Figure G. 8: ROC curve for different classification models (hope)



Online Appendix H: Parliamentary Speeches versus Facebook

In this exercise, we compare the performance of the three different approaches for sentences from different sources (Facebook posts versus legislative speeches). As explained in the main text, there is reason to expect differences between different text types. Legislative speeches are regulated to a great extent (Proksch & Slapin, 2012), while social media circumvent any form of gate keepers. Past research shows that different types of political communication use different emotional language (Widmann, 2021). Therefore, it is reasonable to assume that social media data carries a higher level of emotionally charged language compared to legislative speeches. Indeed, a simple mean comparison reveals that sentences from Facebook contain in average higher amounts of emotional language compared to sentences from parliamentary speeches (see Online Appendix D).

To compare the performance of all three approaches for the two data types, we separated the 10% test data by text source. This resulted in 505 sentences from Facebook and 485 legislative sentences. Then we apply all three approaches to these sub-samples of the test data. Table H.1 shows the results for the different approaches by text type. Table H.2 presents the number of actual and predicted occurrences in an extra table.

The results indicate that the Electra models outperform the other two approaches by far, also in this exercise. This is true for all emotions and for both text types. Furthermore, comparing differences between text sources, it becomes visible that the classification of Facebook sentences achieves higher F1 scores compared to the classification of legislative speeches. This difference is especially striking for the word embeddings and transformer-based approach. This is true for all three tools applied. This finding is in line with previous literature (e.g. Wang et al., 2012) emphasizing the need of high quality test data for machine learning classification.

Table H. 1: Precision, Recall, and F1 scores for the three different approaches by text source

Emotions	Precision	Recall	F1	Emotions	Precision	Recall	F1
ed8 Dictionary							
		Facebook				Parliament	
Anger	0.82	0.49	0.62	Anger	0.84	0.43	0.56
Fear	0.48	0.68	0.56	Fear	0.38	0.63	0.47
Disgust	0.37	0.63	0.47	Disgust	0.21	0.63	0.31
Sadness	0.45	0.58	0.51	Sadness	0.37	0.60	0.45
Joy	0.54	0.63	0.58	Joy	0.37	0.52	0.43
Enthusiasm	0.47	0.51	0.49	Enthusiasm	0.40	0.48	0.43
Pride	0.38	0.53	0.44	Pride	0.24	0.42	0.30
Hope	0.58	0.56	0.57	Hope	0.48	0.49	0.49
Word Embeddings Approach							
		Facebook			Parliament		
Anger	0.79	0.81	0.80	Anger	0.80	0.76	0.78
Fear	0.60	0.54	0.57	Fear	0.64	0.43	0.51
Disgust	0.67	0.51	0.58	Disgust	0.45	0.37	0.41
Sadness	0.72	0.49	0.58	Sadness	0.65	0.34	0.45
Joy	0.77	0.49	0.60	Joy	0.56	0.37	0.44
Enthusiasm	0.65	0.56	0.60	Enthusiasm	0.60	0.43	0.50
Pride	0.57	0.51	0.53	Pride	0.43	0.27	0.33
Hope	0.75	0.60	0.67	Hope	0.63	0.60	0.61
Transformer-based Approach							
		Facebook				Parliament	
Anger	0.84	0.83	0.84	Anger	0.87	0.83	0.85
Fear	0.61	0.73	0.67	Fear	0.58	0.65	0.61
Disgust	0.65	0.69	0.67	Disgust	0.50	0.48	0.49
Sadness	0.68	0.67	0.68	Sadness	0.57	0.44	0.50
Joy	0.70	0.63	0.66	Joy	0.69	0.55	0.61
Enthusiasm	0.67	0.70	0.69	Enthusiasm	0.55	0.65	0.59
Pride	0.68	0.58	0.63	Pride	0.53	0.58	0.56
Hope	0.74	0.84	0.79	Hope	0.61	0.71	0.66

Table H. 2: Actual and Predicted numbers by approach and text source

	Facebook				Parliament			
Emotion	Actual	Predicted by			Actual	Predicted by		
		ed8	Word embeddings	Electra		ed8	Word embeddings	Electra
Anger	252	151	257	250	256	130	243	245
Fear	108	153	97	129	81	134	55	92
Disgust	59	100	45	63	27	82	22	26
Sadness	113	146	76	112	88	142	46	69
Joy	83	94	53	74	60	83	39	48
Enthusiasm	132	142	113	138	88	106	63	104
Pride	91	128	81	78	67	119	42	73
Hope	169	164	136	192	136	139	129	160

Online Appendix I: Dictionary results

I.1: Confusion Matrices for the off-the-shelf dictionaries (test data)

Table I. 1: Confusion Matrix NRC dictionary

		True			True	
	Anger	0	1	**Joy**	0	1
Predicted	0	468	449	0	803	123
	1	14	59	1	44	20
	Fear	0	1	**Enthusiasm**	0	1
Predicted	0	740	153	0	Na	Na
	1	61	36	1	Na	Na
	Disgust	0	1	**Pride**	0	1
Predicted	0	862	80	0	Na	Na
	1	42	6	1	Na	Na
	Sadness	0	1	**Hope**	0	1
Predicted	0	710	164	0	Na	Na
	1	79	37	1	Na	Na

Table I. 2: Confusion Matrix LIWC dictionary

		True			True	
	Anger	0	1	**Joy**	0	1
Predicted	0	441	371	0	Na	Na
	1	41	137	1	Na	Na
	Fear	0	1	**Enthusiasm**	0	1
Predicted	0	751	155	0	Na	Na
	1	50	34	1	Na	Na
	Disgust	0	1	**Pride**	0	1
Predicted	0	Na	Na	0	Na	Na
	1	Na	Na	1	Na	Na
	Sadness	0	1	**Hope**	0	1
Predicted	0	741	156	0	Na	Na
	1	48	45	1	Na	Na

I.2: Dictionary results for the complete test and training data (9898 sentences)

The results presented in Tables I.3 to I.5 Show the performance of the three dictionaries applied to the full dataset (9898 sentences), as a split between training and test data is not necessary for dictionaries. As can be seen, the performance of the individual dictionaries remains similar.

Table I. 3: ed8 Dictionary results for training and test data (9898 sentences)

Emotions	Actual	Predicted	Precision	Recall	F1
Anger	4876	2944	0.77	0.47	0.58
Fear	1985	2951	0.44	0.65	0.52
Disgust	924	1926	0.32	0.66	0.43
Sadness	2012	2860	0.45	0.63	0.53
Joy	1527	1915	0.47	0.59	0.52
Enthusiasm	2122	2487	0.45	0.53	0.49
Pride	2393	1649	0.34	0.49	0.40
Hope	3059	3072	0.54	0.55	0.55

Table I. 4: LIWC Dictionary results for training and test data (9898 sentences)

Emotions	Actual	Predicted	Precision	Recall	F1
Anger	4876	1701	0.72	0.25	0.37
Fear	1985	773	0.41	0.16	0.23
Sadness	2012	1039	0.47	0.24	0.32

Table I. 5: NRC Dictionary results for training and test data (9898 sentences)

Emotions	Actual	Predicted	Precision	Recall	F1
Anger	4876	654	0.76	0.10	0.18
Fear	1985	842	0.32	0.14	0.19
Disgust	924	356	0.20	0.08	0.11
Sadness	2012	1068	0.31	0.16	0.21
Joy	1527	744	0.31	0.15	0.21

I.3: Human judgement against different dictionaries.

In this exercise, we make use of the continuous scale of the ed8 dictionary in order to see whether higher emotional dictionary scores also correlate with higher emotionality as judged by human coders. Human coders, however, had to make a binary decision (emotion associated or not). Therefore, we follow the approach of previous research that argues that inter-annotator agreement/disagreement can be used to determine emotional ambiguity (Andreevskaia & Bergler, 2006; Rauh, 2018; Subasic & Huettner, 2001; Young & Soroka, 2012). The idea is that the more crowd-coders agree on their judgement, the 'clearer' the emotion is present in the respective sentence. Andreesvskaia and Bergler (2006) write that "inter-annotator agreement tends to fall as we proceed from the core of a fuzzy semantic category to its periphery. […] This suggests that inter-annotator disagreement rates can serve as an important source of empirical information about the structural properties of the semantic category" (p. 215).

Thus, human judgment was classified as "clearly emotional" (e.g. clearly angry) when four or five coders coded the sentence as associated with the respective emotion. When two or three coders agreed on one emotion, the human judgment was categorized as "emotional" (e.g. angry) and finally, when none or only one of the coders associated the sentence with the respective emotion the human code was categorized as "not/slightly emotional" (e.g. not/slightly angry). We would expect that the normalized emotional scores significantly increase from the lowest category (not/slightly emotional) to the highest (clearly emotional).

Figure I. 1: Human judgment against different dictionaries (negative emotions)

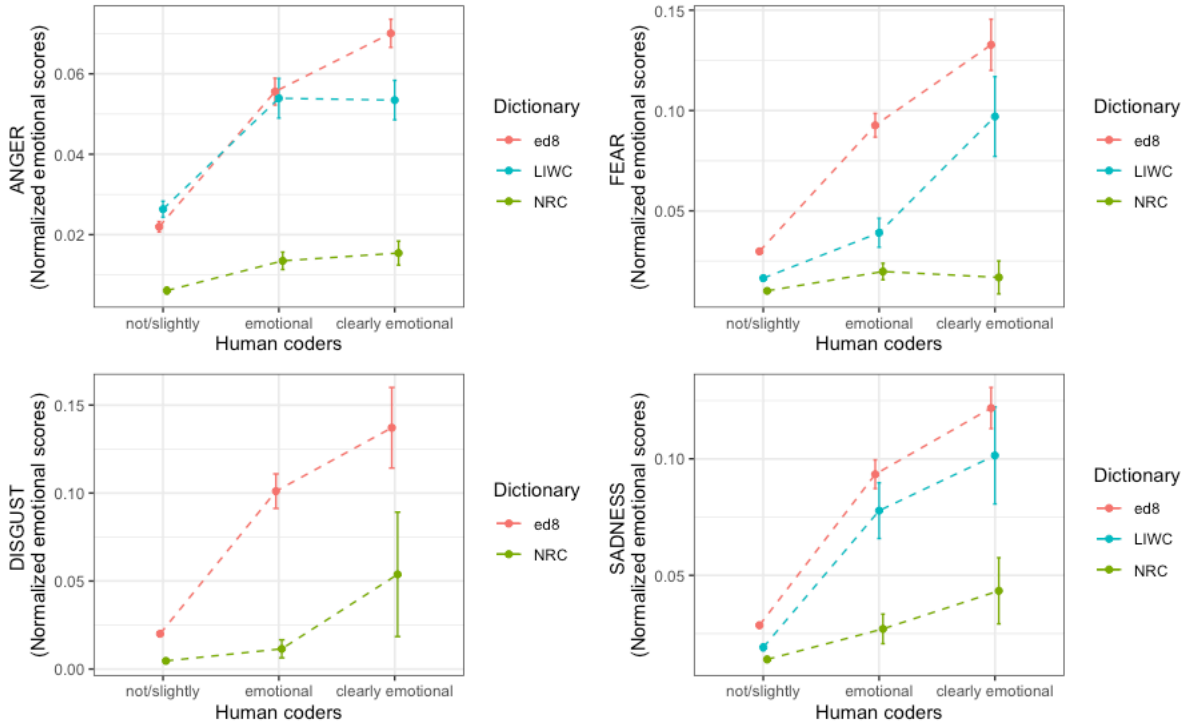
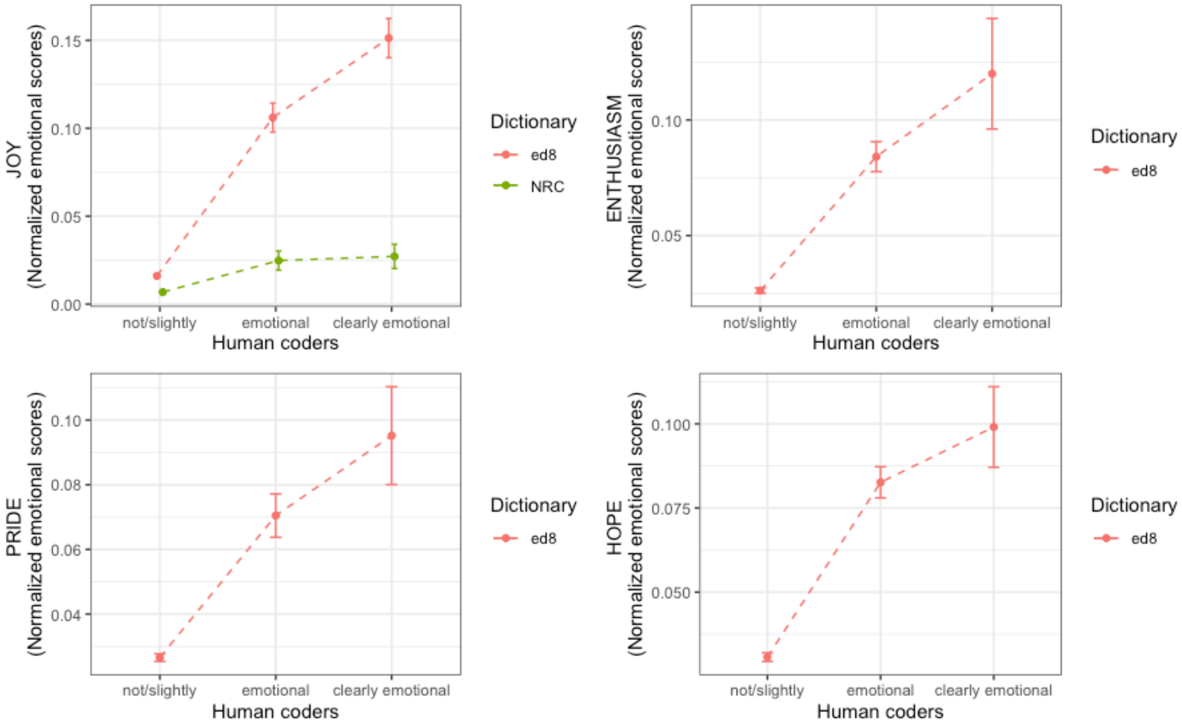


Figure I.1 plots the normalized emotional scores and their 95% confidence intervals for all negative emotions across three categories of human judgment, grouped by dictionary. Figure I.2 shows the plots for all four positive emotions. As can be seen, the NRC EmoLex dictionary shows either no or only a slight increase across the different categories of human judgement. In contrast, the ed8 and the LIWC dictionary exhibit a clearly positive slope across the scale of human judgment. However, the ed8 shows the best performance which discriminates all categories from “not/slightly” to “clearly emotional” in a statistically significant manner with comparably small confidence intervals. The LIWC dictionary shows greater uncertainty and does not discriminate significantly between “angry” and “clearly angry.” Overall, the novel ed8 dictionary applied in this study outperforms widely used off-the-shelf dictionaries in several measures across all comparable emotions. This finding emphasizes the need for domain and language-specific dictionaries.

Figure I. 2: Human judgment against different dictionaries (positive emotions)



Online Appendix J: Replicating the main analysis with randomly sampled crowd-coded sentences

In this exercise we replicate the main analysis using a new dataset of crowd-coded sentences. This new dataset consisted again originally of 10,000 sentences which were selected from the same sample of documents described in Online Appendix B. To make sure that we do not again select the same sentences as in the first dataset, we first excluded the 10,000 sentences from the main analysis. Then we draw a random sample of 5000 sentences from legislative speeches and 5000 sentences from political parties' Facebook posts (same procedure as in the main analysis). This results in 10,000 randomly selected sentences from political communication.

In order to receive human annotation of these sentences, we again used the same crowd-working company called 'Crowdguru'. A random half of the sentences (5000 sentences) were coded again by five individual crowd coders (as in the main analysis), the other half was coded by ten crowd coders each.

The coding process followed the exact same procedure as described in Online Appendix B for the main analysis. Crowd workers were again presented with the same codebook. They also had to conduct a start quiz and their performance was constantly tracked throughout the coding process. In total, six crowd-workers did not pass the 80 percent threshold of correct answers in the start quiz and were therefore dismissed. 101 coders did pass and were therefore eligible to start the coding process. In total, 74 individual coders conducted the coding of the second set of 10,000 sentences.

After dismissing sentences that were coded by at least two or more coders as 'uncodable', 9722 sentences remained. These sentences were then used to calculate precision, recall, and F1 scores for the different approaches. The results of this exercise can be seen in Tables J.1 to J.7. As shown, for all different tools applied, a drop in the performance is noticeable. This drop can be presumably explained through lower levels of emotionality present in the random sample of sentences. Classification models can also 'overfit' on training data and perform very well on test data that is very similar, but poorly on test data that is somewhat different (which might be the case for the randomly sampled sentences).

Nevertheless, the results indicate that the transformer-based model achieves still relatively good F1 scores for most emotions. The gap in F1 scores between the transformer-based approach and the two other tools (word embedding approach and the ed8 dictionary) also further increased. In the main analysis, the transformer-based model's F1 scores were in average 18 points higher per emotion than the ed8 dictionary's scores. Based on the random sample, the Electra model now achieves in average 28.75 points higher F1 scores per emotion. For the word embeddings approach, the average difference in F1 scores per emotion increased from 9.13 to 18.75 in comparison to the transformer-based approach. This finding again exemplifies the strength of the Electra model.

Finally, the performance of the off-the-shelf dictionaries also decreased. For the random sample, the highest F1 score of the NRC dictionary is 0.15 for sadness, the highest F1 score of the LIWC dictionary is 0.06. The lowest F1 score of the NRC dictionary is 0.07 for disgust, while the LIWC dictionary's lowest F1 score is 0.001 for sadness. The finding stresses the differences in performance between customized tools and freely available dictionaries.

Finally, we also replicated the analysis exchanging the locally trained word embeddings with pre-trained word embeddings which achieved comparable results in the main analysis (Bojanowski et al., 2017; Mikolov et al., 2017). The results show that also when applied to a random sample of political sentences, the pre-trained word embeddings combined with a neural network classifier achieve F1 scores in a comparable range as the locally trained embeddings. This is also illustrated in the ROC curves in Figures J.1 to J.8.

Table J. 1: Precision, Recall, and F1 scores for the ed8 dictionary (random sample)

Emotions	Actual	Predicted	Precision	Recall	F1
Anger	4860	948	0.81	0.16	0.26
Fear	1417	1070	0.41	0.31	0.35
Disgust	314	420	0.21	0.28	0.24
Sadness	1233	917	0.34	0.25	0.29
Joy	1204	947	0.37	0.29	0.33
Enthusiasm	2909	1134	0.54	0.21	0.30
Pride	1673	870	0.35	0.18	0.24
Hope	3000	1366	0.53	0.24	0.33

Table J. 2: Precision, Recall, and F1 scores for the word embedding approach (random sample)

Emotions	Actual	Predicted	Precision	Recall	F1
Anger	4860	4556	0.72	0.67	0.70
Fear	1417	941	0.46	0.31	0.37
Disgust	314	265	0.28	0.24	0.26
Sadness	1233	798	0.41	0.27	0.33
Joy	1204	548	0.51	0.23	0.32
Enthusiasm	2909	1251	0.61	0.26	0.37
Pride	1673	955	0.39	0.22	0.28
Hope	3000	2200	0.60	0.44	0.51

Table J. 3: Precision, Recall, and F1 scores for the Electra model (random sample)

Emotions	Actual	Predicted	Precision	Recall	F1
Anger	4860	4437	0.85	0.78	0.81
Fear	1417	1197	0.59	0.50	0.54
Disgust	314	309	0.39	0.39	0.39
Sadness	1233	1104	0.52	0.47	0.49
Joy	1204	903	0.65	0.49	0.56
Enthusiasm	2909	1800	0.74	0.46	0.57
Pride	1673	1318	0.69	0.55	0.61
Hope	3000	3319	0.64	0.71	0.67

Table J. 4: Precision, Recall, and F1 scores for the LIWC dictionary (random sample)

Emotions	Actual	Predicted	Precision	Recall	F1
Anger	4860	687	0.75	0.11	0.19
Fear	1417	415	0.44	0.13	0.20
Sadness	1233	741	0.34	0.20	0.25

Table J. 5: Precision, Recall, and F1 scores for the NRC dictionary (random sample)

Emotions	Actual	Predicted	Precision	Recall	F1
Anger	4860	378	0.73	0.06	0.11
Fear	1417	561	0.23	0.09	0.13
Disgust	314	201	0.09	0.06	0.07
Sadness	1233	791	0.19	0.12	0.15
Joy	1204	569	0.21	0.10	0.13

Table J. 6: Precision, Recall, and F1 scores for pre-trained word embeddings (Bojanowski et al. 2017, random sample)

Emotions	Actual	Predicted	Precision	Recall	F1
Anger	4860	4925	0.68	0.69	0.69
Fear	1417	1271	0.43	0.38	0.40
Disgust	314	391	0.31	0.38	0.34
Sadness	1233	761	0.46	0.28	0.35
Joy	1204	507	0.50	0.21	0.29
Enthusiasm	2909	1518	0.58	0.30	0.40
Pride	1673	665	0.44	0.17	0.25
Hope	3000	2312	0.57	0.44	0.49

Table J. 7: Precision, Recall, and F1 scores for pre-trained word embeddings (Mikolov et al. 2017, random sample)

Emotions	Actual	Predicted	Precision	Recall	F1
Anger	4860	4539	0.72	0.67	0.69
Fear	1417	930	0.48	0.31	0.38
Disgust	314	225	0.32	0.23	0.26
Sadness	1233	1195	0.38	0.37	0.38
Joy	1204	414	0.55	0.19	0.28
Enthusiasm	2909	862	0.66	0.19	0.30
Pride	1673	721	0.44	0.19	0.26
Hope	3000	2139	0.60	0.43	0.50

Table J. 8: Confusion Matrix for the ed8 dictionary (random sample)

		True			True	
	Anger	0	1	**Joy**	0	1
Predicted	0	4679	4095	0	7924	851
	1	183	765	1	594	353
	Fear	0	1	**Enthusiasm**	0	1
Predicted	0	7671	981	0	6292	2296
	1	634	436	1	521	613
	Disgust	0	1	**Pride**	0	1
Predicted	0	9076	226	0	7480	1372
	1	332	88	1	569	301
	Sadness	0	1	**Hope**	0	1
Predicted	0	7881	924	0	6083	2273
	1	608	309	1	639	727

Table J. 9: Confusion Matrix for the word embedding approach (random sample)

		True			True	
	Anger	0	1	**Joy**	0	1
Predicted	0	3583	1583	0	8252	922
	1	1279	3277	1	266	282
	Fear	0	1	**Enthusiasm**	0	1
Predicted	0	7797	984	0	6326	2145
	1	508	433	1	487	764
	Disgust	0	1	**Pride**	0	1
Predicted	0	9217	240	0	7467	1300
	1	191	74	1	582	373
	Sadness	0	1	**Hope**	0	1
Predicted	0	8022	902	0	5843	1679
	1	467	331	1	879	1321

Table J. 10: Confusion Matrix for the Electra model (random sample)

		True			True	
	Anger	0	1	**Joy**	0	1
Predicted	0	4206	1079	0	8203	616
	1	656	3781	1	315	588
	Fear	0	1	**Enthusiasm**	0	1
Predicted	0	7810	715	0	6346	1576
	1	495	702	1	467	1333
	Disgust	0	1	**Pride**	0	1
Predicted	0	9221	192	0	7644	760
	1	187	122	1	405	913
	Sadness	0	1	**Hope**	0	1
Predicted	0	7959	659	0	5525	878
	1	530	574	1	1197	2122

Table J. 11: Confusion Matrix for the NRC dictionary (random sample)

		True			True	
	Anger	0	1	**Joy**	0	1
Predicted	0	4761	4583	0	8068	1085
	1	101	277	1	450	119
	Fear	0	1	**Enthusiasm**	0	1
Predicted	0	7874	1287	0	Na	Na
	1	431	130	1	Na	Na
	Disgust	0	1	**Pride**	0	1
Predicted	0	9226	295	0	Na	Na
	1	182	19	1	Na	Na
	Sadness	0	1	**Hope**	0	1
Predicted	0	7846	1085	0	Na	Na
	1	643	148	1	Na	Na

Table J. 12: Confusion Matrix for the LIWC dictionary (random sample)

		True			True	
	Anger	0	1	**Joy**	0	1
Predicted	0	4692	4343	0	Na	Na
	1	170	517	1	Na	Na
	Fear	0	1	**Enthusiasm**	0	1
Predicted	0	8074	1233	0	Na	Na
	1	231	184	1	Na	Na
	Disgust	0	1	**Pride**	0	1
Predicted	0	Na	Na	0	Na	Na
	1	Na	Na	1	Na	Na
	Sadness	0	1	**Hope**	0	1
Predicted	0	7997	984	0	Na	Na
	1	492	249	1	Na	Na

Table J. 13: Confusion Matrix for pre-trained word embeddings (Bojanowski et al. 2017, random sample)

		True			True	
	Anger	0	1	**Joy**	0	1
Predicted	0	3302	1495	0	8263	952
	1	1560	3365	1	255	252
	Fear	0	1	**Enthusiasm**	0	1
Predicted	0	7575	876	0	6182	2022
	1	730	541	1	631	887
	Disgust	0	1	**Pride**	0	1
Predicted	0	9137	194	0	7676	1381
	1	271	120	1	373	292
	Sadness	0	1	**Hope**	0	1
Predicted	0	8076	885	0	5718	1692
	1	413	348	1	1004	1308

Table J. 14: Confusion Matrix for pre-trained word embeddings (Mikolov et al. 2017, random sample)

		True			True	
	Anger	0	1	**Joy**	0	1
Predicted	0	3577	1606	0	8332	976
	1	1285	3254	1	186	228
	Fear	0	1	**Enthusiasm**	0	1
Predicted	0	7820	972	0	6518	2342
	1	485	445	1	295	567
	Disgust	0	1	**Pride**	0	1
Predicted	0	9254	243	0	7644	1357
	1	154	71	1	405	316
	Sadness	0	1	**Hope**	0	1
Predicted	0	7754	773	0	5867	1716
	1	735	460	1	855	1284

Figure J. 1: ROC curve for different classification models (anger)

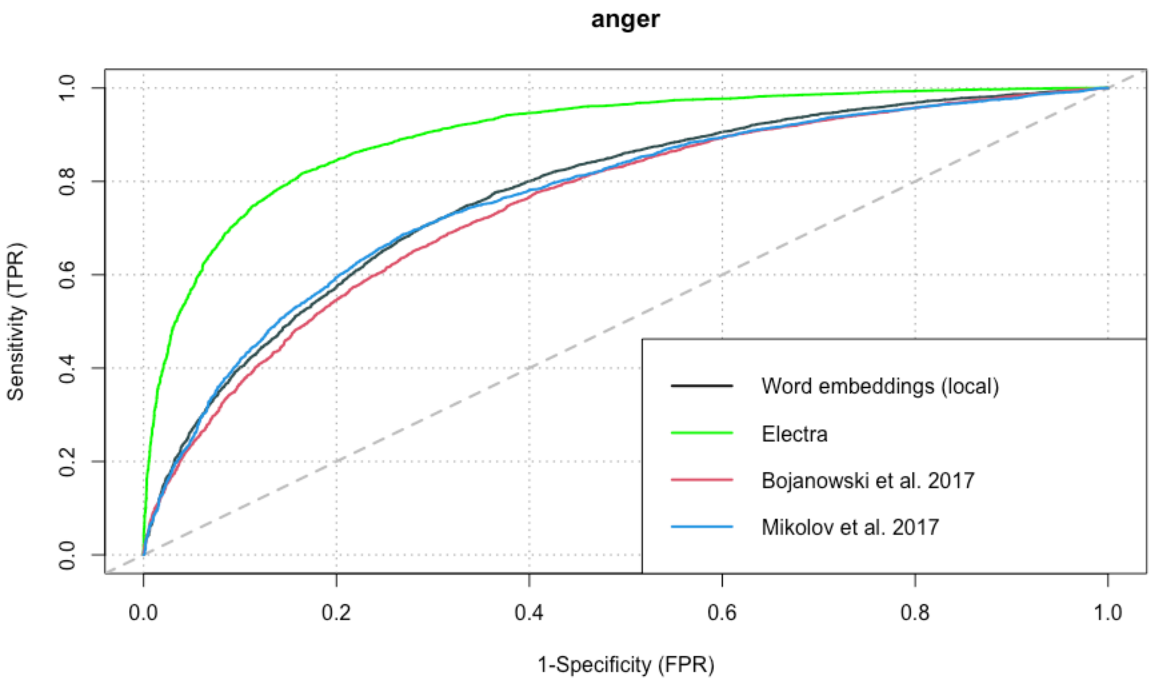


Figure J. 2: ROC curve for different classification models (fear)

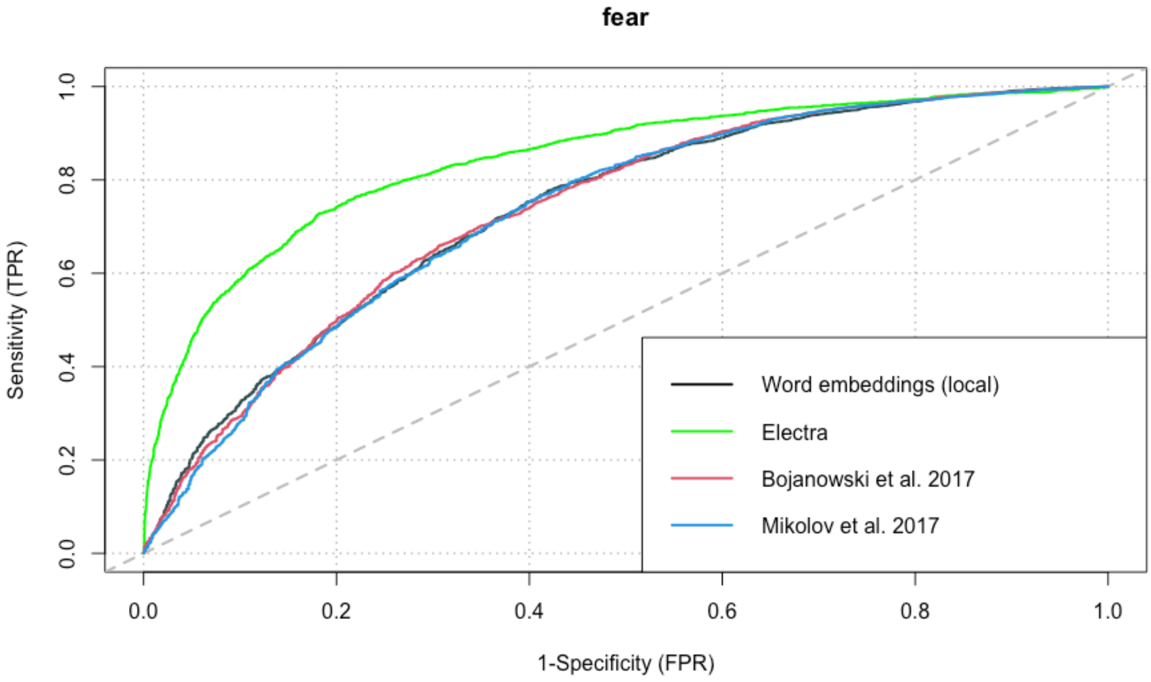


Figure J. 3: ROC curve for different classification models (disgust)

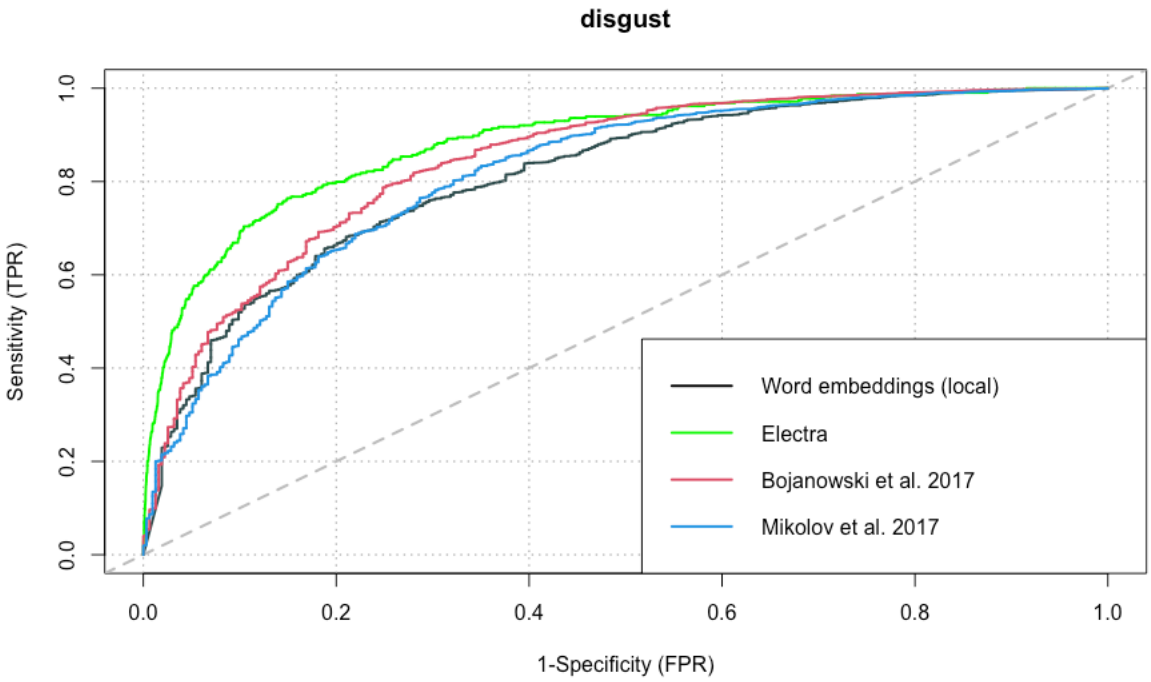


Figure J. 4: ROC curve for different classification models (sadness)

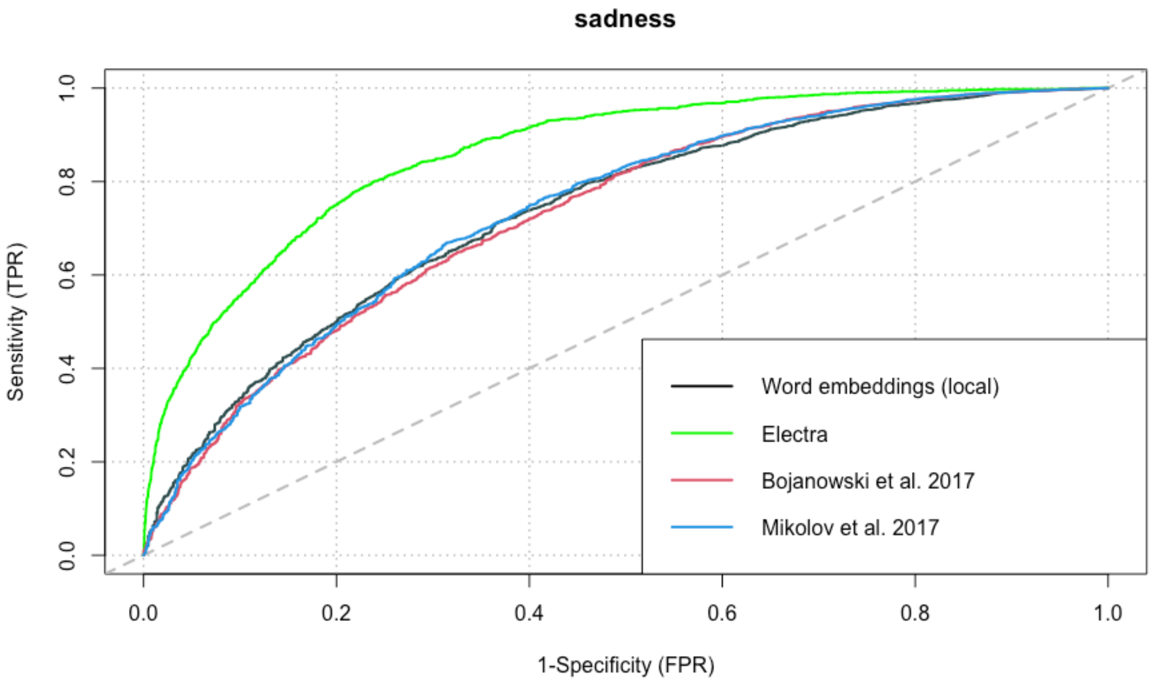


Figure J. 5: ROC curve for different classification models (joy)

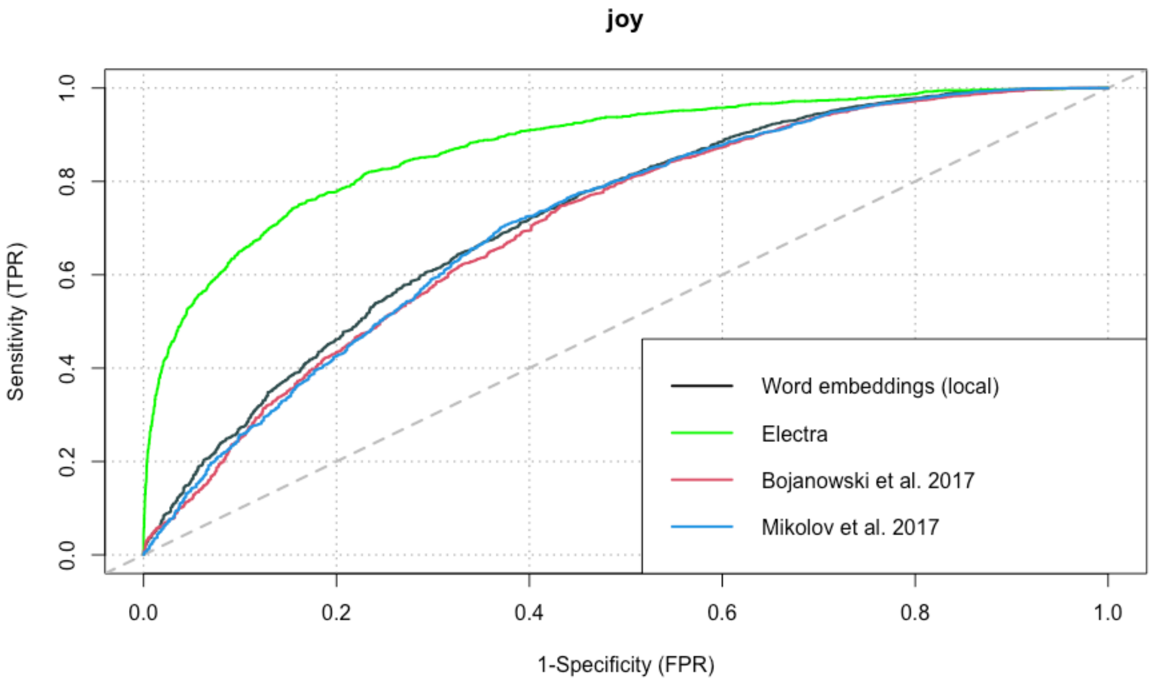


Figure J. 6: ROC curve for different classification models (enthusiasm)

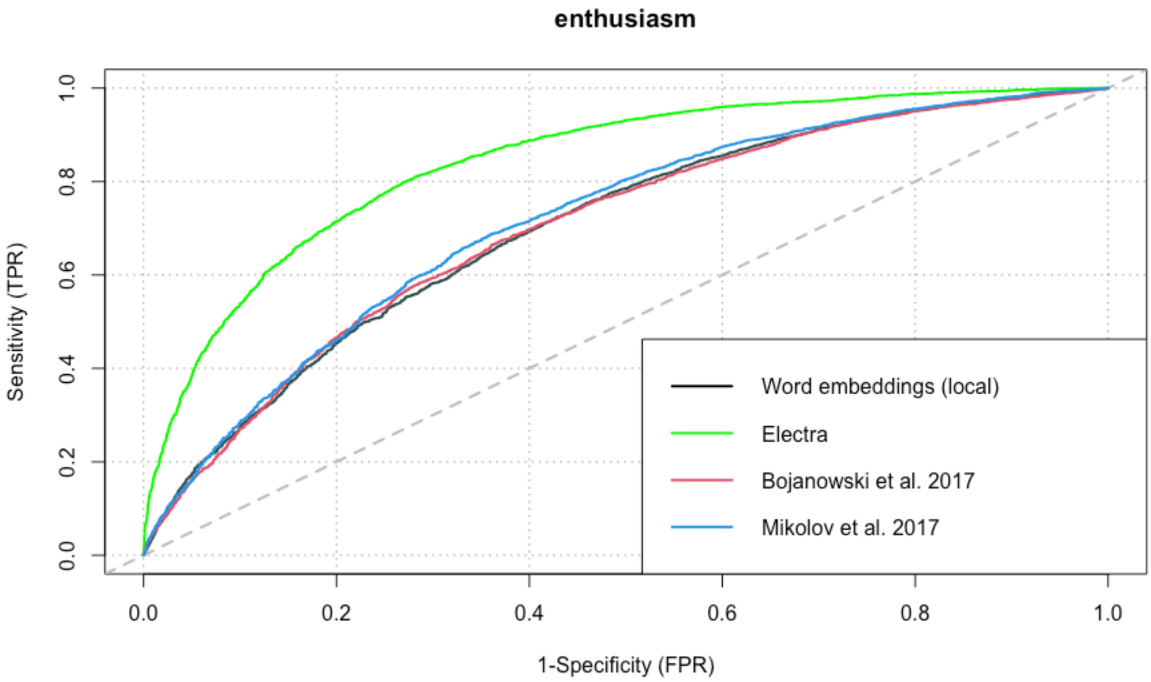


Figure J. 7: ROC curve for different classification models (pride)

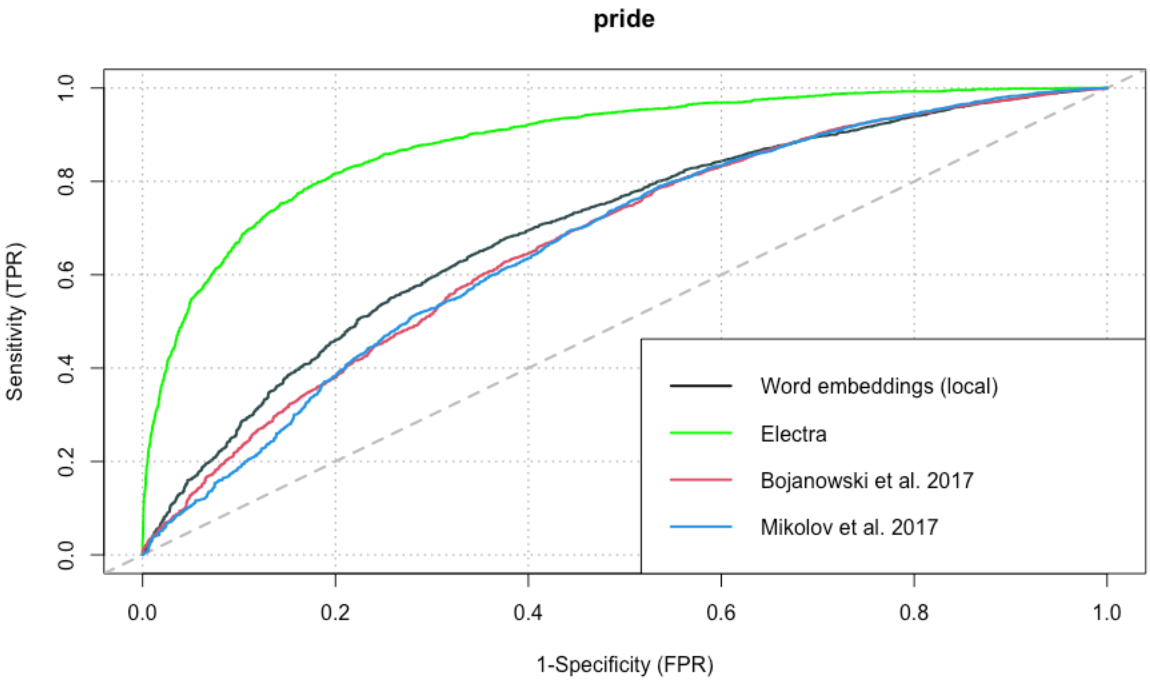
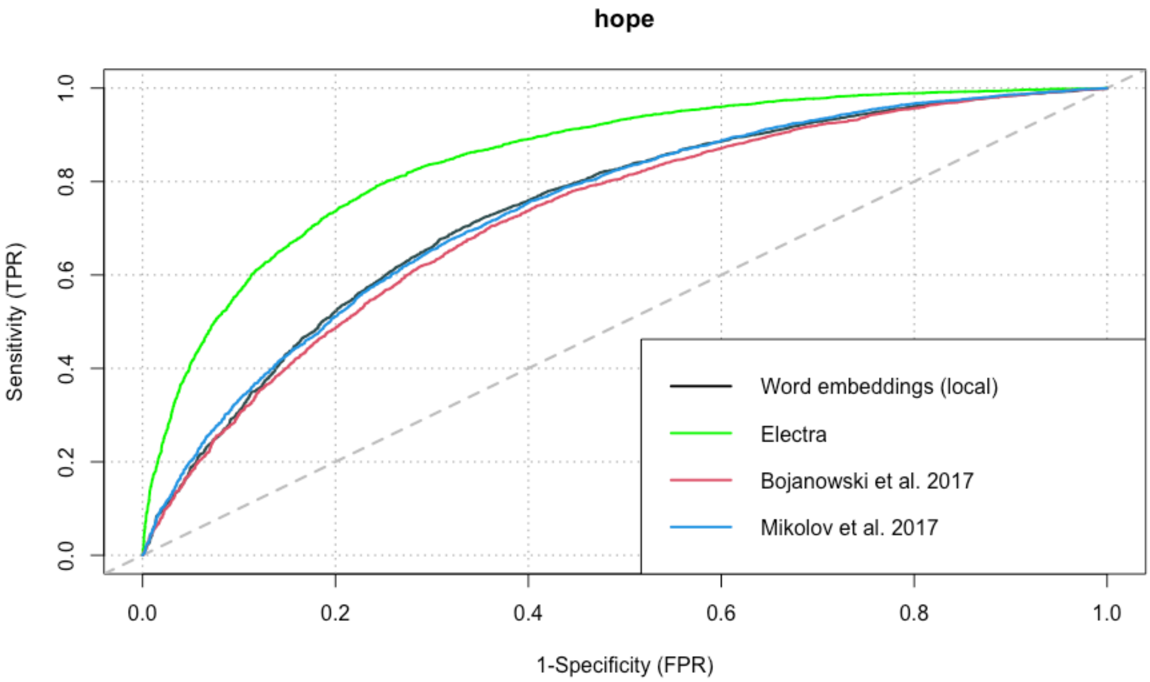


Figure J. 8: ROC curve for different classification models (hope)



Online Appendix K: Bootstrapping exercise

In this exercise we take the 5000 sentences from the second crowd-coded dataset which have been coded by 10 individuals each. Then we draw on random subsamples from these 5000 sentences to estimate F1 scores as a function of crowd coders per sentence. We do so by bootstrapping 1000 sets of subsamples with replacement for each *n* ranging from *n* = 1 to *n* = 10 coders per sentence. To be precise, we randomly select *n* judgements from the overall 10 judgements per sentence. Then we turn these randomly selected judgements again into a binary variable (as in the main analysis, a sentence is coded as emotional as long as one human judgement coded it as such). Subsequently, we calculate the F1 score for each tool. We replicate this process 1,000 times for each *n*.

The results are displayed below in Figure K.1 and K.2. As one would assume, the F1 scores increase with increasing number of coders. The slopes are relatively steep in the beginning as the judgements per sentence increase from 1 to 2, from 2 to 3, and from 3 to 4. However, the slopes flatten with increasing numbers of coders, and even turns negative at some point for some emotions (joy and enthusiasm). This makes sense since the F1 score is calculated based on precision and recall. With increasing numbers of coders, the precision scores of the different approaches increase (see Tables K.1 to K.8) because the increasing number of data points help the models to better distinguish between the classes. This confirms the validity of the crowd-sourcing approach, which is in line with previous studies (e.g. Benoit et al., 2016). Yet, at the same time the recall values decrease because more coders mean more sentences judged as emotional. If there are too many judgements per sentences, it will automatically increase the amount of 'true emotional sentences' which then results in a decreasing recall value. This means the model's ability to detect the 'more valid' emotions decreases. Hence, if we would increase the number of coders even more beyond 10 one should expect that the curves begin to drop again (as they already do for joy and enthusiasm). Therefore, 5 judgements per sentence might be appropriate choice especially since in average we reach already 92.8 percent of the F1 scores we achieve when we obtain 10 judgements per sentence (based on the Electra model). Additionally, 5 judgements provide a decent trade-off between precision and recall. A low recall impairs the external validity because a model would detect only a limited number of the relevant items. From a cost-benefit and a precision-recall perspective, we thus argue that 5 judgements per sentence should be sufficient.

Figure K. 1: F1 scores as a function of the number of coders by different approach (negative emotions)

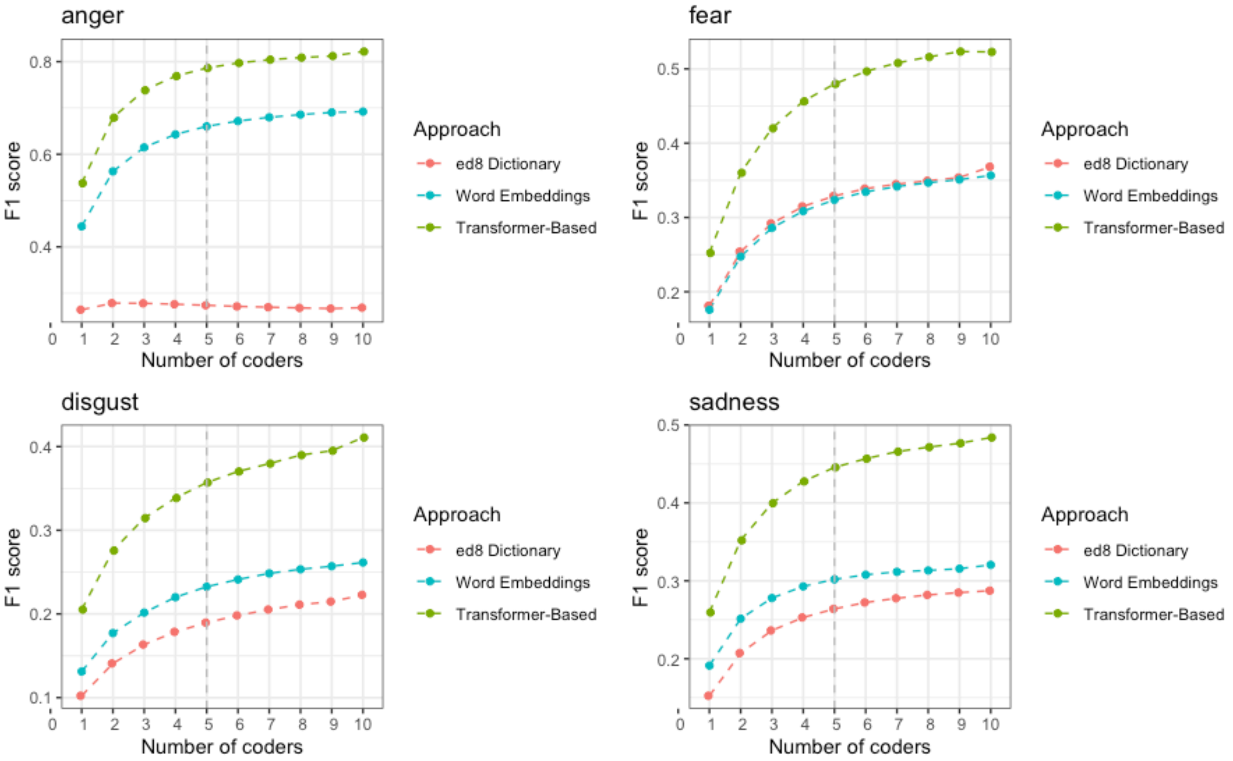


Figure K. 2: F1 scores as a function of the number of coders by different approach (positive emotions)

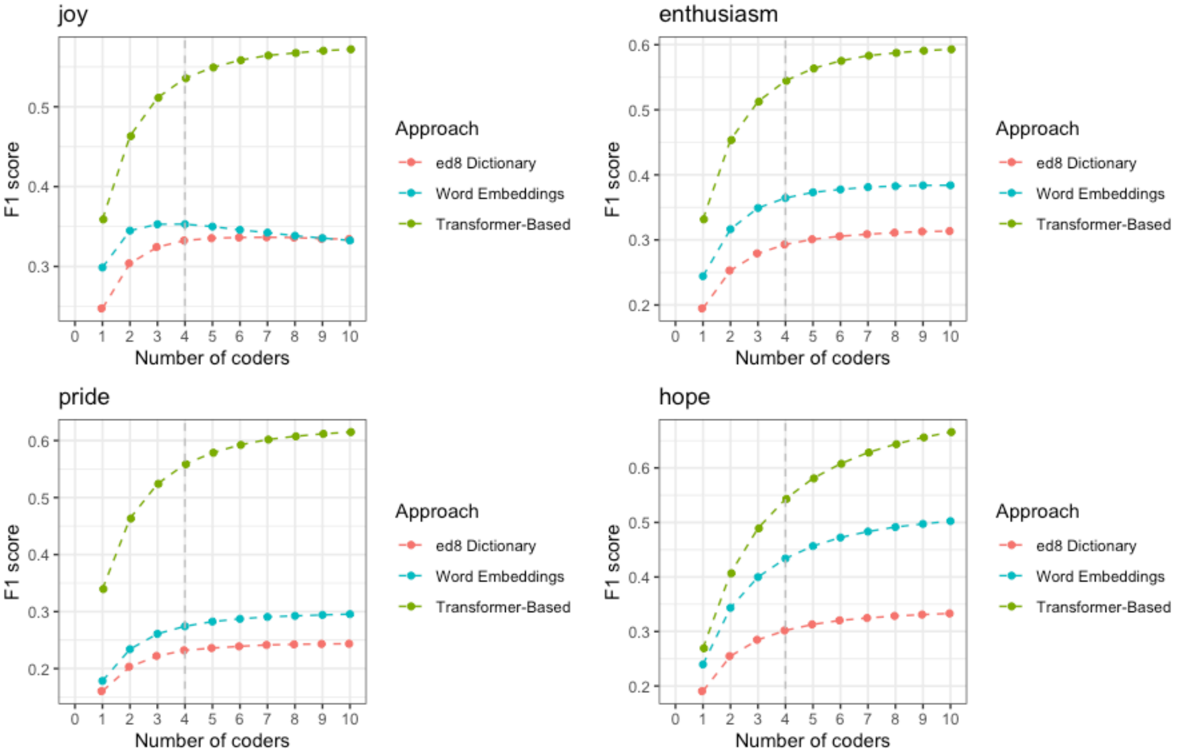


Table K. 1: Precision, Recall, and F1 score by the number of coders per sentence (anger)

Anger	Precision	Recall	F1 scores	Precision	Recall	F1 scores	Precision	Recall	F1 scores
Number of coders		ed8 Dictionary			Word Embeddings			Electra	
1	0,40	0,20	0,26	0,32	0,74	0,44	0,39	0,87	0,54
2	0,55	0,19	0,28	0,46	0,72	0,56	0,57	0,85	0,68
3	0,63	0,18	0,28	0,54	0,71	0,61	0,66	0,83	0,74
4	0,68	0,17	0,28	0,59	0,70	0,64	0,72	0,82	0,77
5	0,71	0,17	0,27	0,63	0,69	0,66	0,76	0,81	0,79
6	0,73	0,17	0,27	0,66	0,69	0,67	0,79	0,80	0,80
7	0,75	0,16	0,27	0,68	0,69	0,68	0,81	0,79	0,80
8	0,76	0,16	0,27	0,69	0,68	0,69	0,83	0,79	0,81
9	0,77	0,16	0,27	0,70	0,68	0,69	0,84	0,78	0,81
10	0,78	0,16	0,27	0,71	0,68	0,69	0,86	0,78	0,81

Table K. 2: Precision, Recall, and F1 score by the number of coders per sentence (fear)

Fear	Precision	Recall	F1 scores	Precision	Recall	F1 scores	Precision	Recall	F1 scores
Number of coders		ed8 Dictionary			Word Embeddings			Electra	
1	0,12	0,40	0,18	0,12	0,35	0,18	0,16	0,59	0,25
2	0,19	0,38	0,25	0,20	0,34	0,25	0,26	0,57	0,36
3	0,24	0,37	0,29	0,25	0,33	0,29	0,34	0,56	0,42
4	0,28	0,36	0,31	0,29	0,33	0,31	0,39	0,55	0,46
5	0,31	0,35	0,33	0,33	0,32	0,32	0,43	0,54	0,48
6	0,34	0,34	0,34	0,35	0,32	0,33	0,47	0,53	0,50
7	0,35	0,34	0,35	0,37	0,31	0,34	0,50	0,52	0,51
8	0,37	0,33	0,35	0,39	0,31	0,35	0,52	0,51	0,52
9	0,38	0,33	0,35	0,41	0,31	0,35	0,54	0,51	0,52
10	0,41	0,33	0,37	0,42	0,31	0,36	0,56	0,50	0,53

Table K. 3: Precision, Recall, and F1 score by the number of coders per sentence (disgust)

Disgust	Precision	Recall	F1 scores	Precision	Recall	F1 scores	Precision	Recall	F1 scores
Number of coders	ed8 Dictionary			Word Embeddings			Electra		
1	0,06	0,37	0,10	0,08	0,32	0,13	0,13	0,54	0,21
2	0,09	0,34	0,14	0,13	0,29	0,18	0,19	0,49	0,27
3	0,11	0,32	0,16	0,16	0,28	0,20	0,24	0,46	0,31
4	0,13	0,30	0,18	0,18	0,27	0,22	0,28	0,44	0,34
5	0,14	0,29	0,19	0,21	0,26	0,23	0,31	0,43	0,36
6	0,15	0,29	0,20	0,22	0,26	0,24	0,33	0,42	0,37
7	0,16	0,28	0,21	0,24	0,26	0,25	0,35	0,41	0,38
8	0,17	0,28	0,21	0,26	0,25	0,25	0,37	0,41	0,39
9	0,18	0,27	0,22	0,27	0,25	0,26	0,39	0,40	0,39
10	0,18	0,27	0,22	0,28	0,25	0,26	0,40	0,40	0,40

Table K. 4: Precision, Recall, and F1 score by the number of coders per sentence (sadness)

Sadness	Precision	Recall	F1 scores	Precision	Recall	F1 scores	Precision	Recall	F1 scores
Number of coders	ed8 Dictionary			Word Embeddings			Electra		
1	0,10	0,32	0,15	0,13	0,34	0,19	0,16	0,62	0,26
2	0,16	0,30	0,21	0,21	0,32	0,25	0,25	0,58	0,35
3	0,20	0,29	0,24	0,26	0,30	0,28	0,32	0,55	0,40
4	0,23	0,29	0,25	0,30	0,29	0,29	0,36	0,53	0,43
5	0,25	0,28	0,26	0,32	0,28	0,30	0,39	0,51	0,45
6	0,27	0,27	0,27	0,35	0,28	0,31	0,42	0,50	0,46
7	0,29	0,27	0,28	0,37	0,27	0,31	0,44	0,49	0,47
8	0,30	0,27	0,28	0,38	0,27	0,31	0,46	0,48	0,47
9	0,31	0,26	0,29	0,39	0,26	0,32	0,48	0,48	0,48
10	0,32	0,26	0,29	0,41	0,26	0,32	0,49	0,47	0,48

Table K. 5: Precision, Recall, and F1 score by the number of coders per sentence (joy)

Joy	Precision	Recall	F1 scores	Precision	Recall	F1 scores	Precision	Recall	F1 scores
Number of coders	ed8 Dictionary			Word Embeddings			Electra		
1	0,17	0,46	0,25	0,24	0,39	0,30	0,24	0,68	0,36
2	0,24	0,41	0,30	0,34	0,35	0,34	0,37	0,63	0,46
3	0,28	0,38	0,32	0,40	0,32	0,35	0,44	0,60	0,51
4	0,31	0,36	0,33	0,44	0,30	0,35	0,50	0,58	0,54
5	0,33	0,34	0,34	0,46	0,28	0,35	0,54	0,56	0,55
6	0,34	0,33	0,34	0,48	0,27	0,35	0,57	0,55	0,56
7	0,36	0,32	0,34	0,49	0,26	0,34	0,59	0,54	0,56
8	0,37	0,31	0,34	0,51	0,25	0,34	0,61	0,53	0,57
9	0,37	0,30	0,33	0,52	0,25	0,34	0,63	0,52	0,57
10	0,38	0,30	0,33	0,52	0,24	0,33	0,64	0,52	0,57

Table K. 6: Precision, Recall, and F1 score by the number of coders per sentence (enthusiasm)

Enthusiasm	Precision	Recall	F1 scores	Precision	Recall	F1 scores	Precision	Recall	F1 scores
Number of coders	ed8 Dictionary			Word Embeddings			Electra		
1	0,15	0,27	0,19	0,19	0,35	0,24	0,23	0,61	0,33
2	0,25	0,26	0,25	0,30	0,33	0,32	0,37	0,59	0,45
3	0,32	0,25	0,28	0,38	0,32	0,35	0,47	0,57	0,51
4	0,37	0,24	0,29	0,44	0,31	0,36	0,54	0,55	0,54
5	0,40	0,24	0,30	0,49	0,30	0,37	0,59	0,54	0,56
6	0,43	0,24	0,31	0,52	0,30	0,38	0,63	0,53	0,58
7	0,46	0,23	0,31	0,55	0,29	0,38	0,67	0,52	0,58
8	0,48	0,23	0,31	0,57	0,29	0,38	0,69	0,51	0,59
9	0,50	0,23	0,31	0,59	0,28	0,38	0,71	0,50	0,59
10	0,51	0,23	0,31	0,60	0,28	0,38	0,73	0,50	0,59

Table K. 7: Precision, Recall, and F1 score by the number of coders per sentence (pride)

Pride	Precision	Recall	F1 scores	Precision	Recall	F1 scores	Precision	Recall	F1 scores
Number of coders	ed8 Dictionary			Word Embeddings			Electra		
1	0,12	0,25	0,16	0,13	0,28	0,18	0,23	0,68	0,34
2	0,18	0,23	0,20	0,21	0,27	0,23	0,36	0,66	0,46
3	0,23	0,22	0,22	0,26	0,27	0,26	0,45	0,64	0,52
4	0,26	0,21	0,23	0,29	0,26	0,27	0,51	0,62	0,56
5	0,28	0,21	0,24	0,32	0,25	0,28	0,55	0,61	0,58
6	0,29	0,20	0,24	0,34	0,25	0,29	0,59	0,60	0,59
7	0,31	0,20	0,24	0,36	0,25	0,29	0,62	0,59	0,60
8	0,32	0,20	0,24	0,37	0,24	0,29	0,64	0,58	0,61
9	0,33	0,19	0,24	0,38	0,24	0,29	0,66	0,57	0,61
10	0,33	0,19	0,24	0,39	0,24	0,30	0,67	0,56	0,61

Table K. 8: Precision, Recall, and F1 score by the number of coders per sentence (hope)

Hope	Precision	Recall	F1 scores	Precision	Recall	F1 scores	Precision	Recall	F1 scores
Number of coders	ed8 Dictionary			Word Embeddings			Electra		
1	0,14	0,29	0,19	0,16	0,50	0,24	0,16	0,78	0,27
2	0,24	0,28	0,25	0,27	0,48	0,34	0,28	0,77	0,41
3	0,30	0,27	0,28	0,34	0,48	0,40	0,36	0,76	0,49
4	0,35	0,26	0,30	0,40	0,47	0,43	0,42	0,75	0,54
5	0,39	0,26	0,31	0,45	0,46	0,46	0,47	0,75	0,58
6	0,42	0,26	0,32	0,49	0,46	0,47	0,51	0,74	0,61
7	0,45	0,26	0,32	0,52	0,45	0,48	0,55	0,74	0,63
8	0,47	0,25	0,33	0,54	0,45	0,49	0,57	0,73	0,64
9	0,49	0,25	0,33	0,56	0,45	0,50	0,60	0,73	0,66
10	0,50	0,25	0,33	0,58	0,44	0,50	0,62	0,73	0,67

Online Appendix L: Comparing normalized emotional scores with and without stop words

In this exercise, we test whether the exclusion of stop words from the calculation of the normalized emotional scores of the ed8 dictionary does bias our results. Excluding stop words could overestimate the relevance of emotion words compared to the proportion of all words. This means that the ed8 dictionary could potentially signal strong emotional language which is not necessarily perceived by humans as such. However, since the exclusion of stop words can only influence the continuous emotional score (proportion of emotional words per sentence) and not the binary variable which we use to calculate the F1 scores in the main analysis, we only replicate the comparison of the continuous scale to human judgement.

The results of this exercise can be seen in the figures below. If the exclusion of stop words would have introduced a significant bias, one could expect that the two lines in Figure L.1 and L.2 would not be parallel. Overestimating emotion words could lead to high dictionary scores even though human coders would not perceive these sentences as overly emotional. Hence, the blue line should not be able to discriminate significantly between the different categories but instead be closer to a horizontal line or even exhibit a negative slope.

Figure L. 1: Human judgment against different emotional scores (negative emotions)

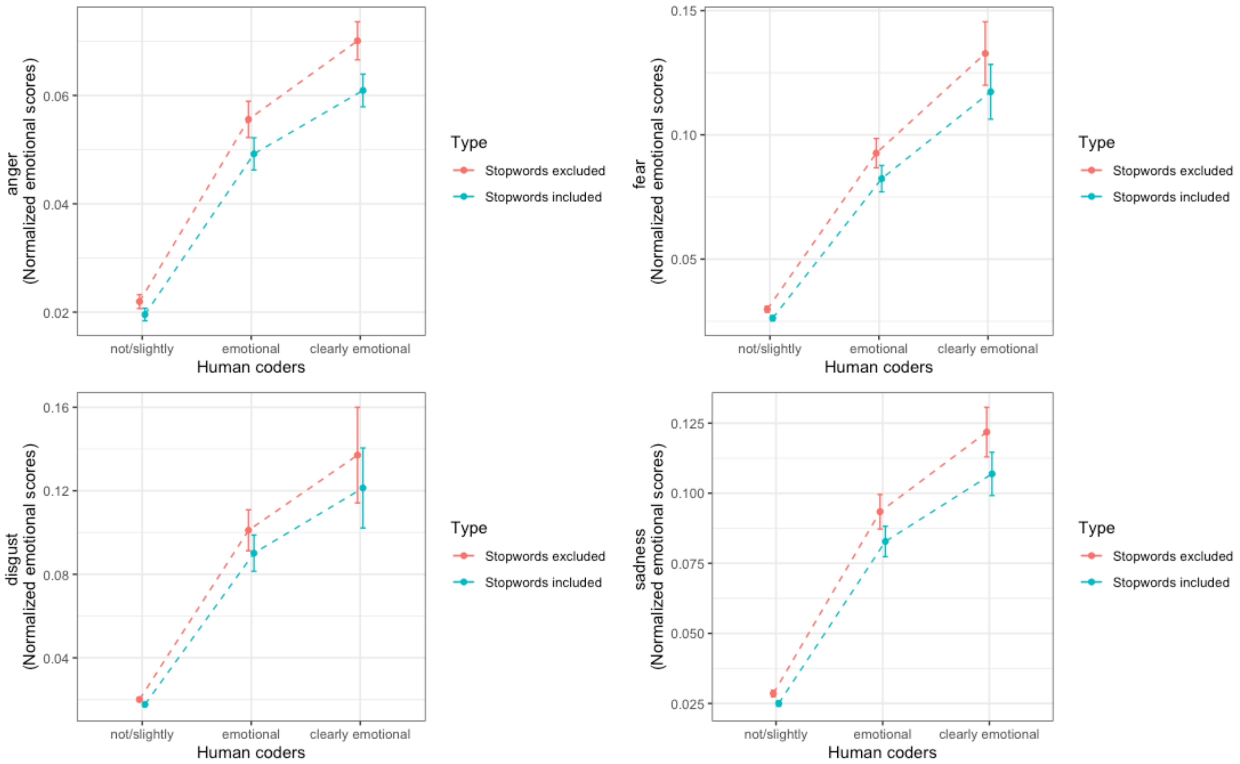
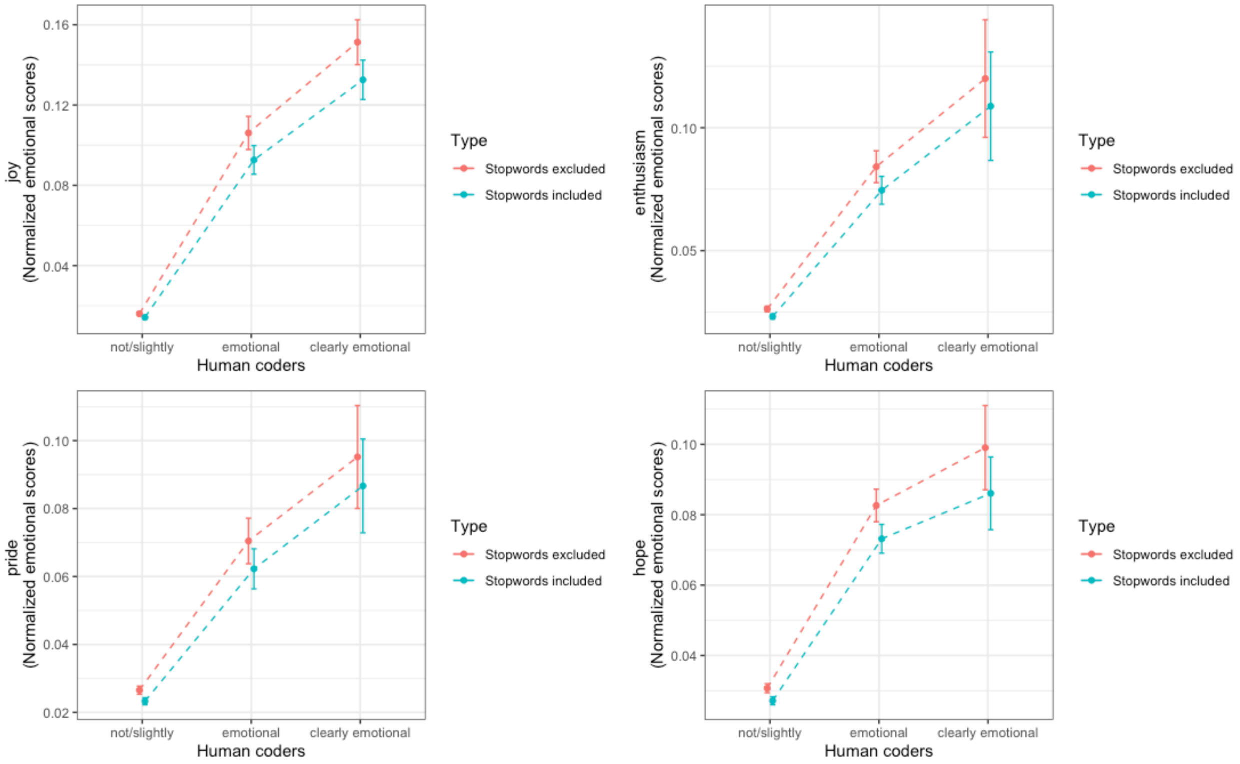


Figure L. 2: Figure K. 4: Human judgment against different emotional scores (positive emotions)



Moreover, an additional analysis shows that there is no strong correlation between stop words and specific emotions (as judged by the crowd coders).

Table L. 1: Pearson correlation between stop words and emotions

	Stop words	Anger	Fear	Disgust	Sadness	Joy	Enthusiasm	Pride	Hope
Stop words	1,000								
Anger	0,078	1,000							
Fear	0,024	0,286	1,000						
Disgust	0,009	0,253	0,232	1,000					
Sadness	0,030	0,281	0,299	0,321	1,000				
Joy	-0,013	-0,322	-0,182	-0,123	-0,182	1,000			
Enthusiasm	-0,012	-0,328	-0,177	-0,143	-0,204	0,069	1,000		
Pride	0,019	-0,328	-0,196	-0,132	-0,178	0,362	0,264	1,000	
Hope	0,022	-0,375	-0,215	-0,188	-0,231	0,096	0,434	0,200	1,000

Online Appendix M: Application Example: Analyzing party press releases from Germany

To increase the external validity of the tools created in this study, we conduct a short case study including two sets of hypotheses. We test these hypotheses relying on the transformer-based Electra model which we chose due to the highest performance in measuring discrete emotional appeals. We apply the model to a large dataset of press releases from six political parties in Germany. In total, the case study analyzes 12,580 press releases

Firstly, we aim at analyzing how different party groups (government, opposition, and radical parties) use emotions differently. Secondly, we aim at analyzing which specific emotions parties appeal to during election campaigns, in comparison to routine times. If it is indeed true that parties appeal to emotions strategically, as recent literature suggests (Crabtree et al., 2020; Kosmidis et al., 2019), our newly created and validated tools should be able to capture subtle differences between party groups and between routine and campaign periods.

The German party system consists momentarily of six major parties represented in the national parliament. The social democrats (SPD) and the Christian democrats (CDU/CSU) currently form the government coalition. Mainstream opposition parties are the Green Party and the liberals (FDP). The Left (Die Linke) and the 'Alternative for Germany' (AfD) are considered to be radical populist parties (see Rooduijn et al., 2019). Furthermore, the last German national elections took place on September 24, 2017. The research period includes press releases from the beginning of 2016 until the end of 2018.

The typical communication strategies of government, opposition, and radical parties should produce systematic differences in the usage of emotional appeals which we can use to strengthen the external validity of our tools. The same is to be expected about differences between routine and campaign periods. We therefore put forward the following hypotheses.

M.1 Hypotheses

The first two hypotheses concern differences in emotional appeals between party groups. A major task of opposition parties is to criticize the government's policies and work. This criticism is often particularly harsh from radical parties on the fringes of the political spectrum. These parties aim at attacking the competence of mainstream competitors, such as government actors, by bringing up new divisive issues and using anti-establishment rhetoric (De Vries & Hobolt, 2020).

This rhetoric often aims at portraying the elites and the political establishment as corrupt and morally bankrupt. Populist radical parties in particular have been found to make use of such communication strategies in order to attack and blame the elites and other groups for the grievances of society (Mudde, 2004). Prior research found that they often rely on negative emotional appeals to do so (Hameleers et al., 2016). It is therefore reasonable to expect that they use higher levels of negative emotional appeals in their communication.

Mainstream actors in the center of the political spectrum, on the other hand, should be interested in framing the state of the political world in positive terms by emphasizing the achievements of the political establishment. We therefore expect mainstream parties to be more positive and less negative than radical parties. This should be true for both mainstream opposition and government parties. After all, all mainstream parties routinely switch from government to opposition and are therefore considered to be in a winning position (Hobolt & de Vries, 2015). However, we still expect differences between mainstream opposition and incumbent parties. This is due to the fact that opposition parties need to distinguish themselves from government parties in order to gain electoral benefits. In order to emphasize differences between themselves and the political opponents, parties can use emotional language (Kosmidis et al., 2019). Mainstream opposition parties are therefore expected to be in the middle ground: less negative than radical parties but more negative than incumbents. The same is expected for positive emotions: mainstream opposition parties should be less positive than government parties but more positive than radical challengers on the fringes of the political spectrum.

Finally, government parties have the most incentives to use positive emotional appeals. This is based on the idea that individuals reward the incumbent party when they perceive the current situation as good. Parties should therefore try to shape those perceptions through their political communication (Crabtree et al., 2020). Parties could do this for example by using retrospective positive emotions, such as pride or joy, which portrays their past achievements in positive terms. They could also create a more positive outlook for the future, relying on prospective positive emotions such as enthusiasm and hope, in order to connect their past achievements with the future to come. These strategies can be successful in increasing voter support for government as research shows that emotions can influence the way citizens' process information (Utych, 2018) and judge incumbents (Healy et al., 2010). Government parties, who are perceived as responsible for the state of affairs, are therefore expected to use the highest level of positive emotions and the lowest level of negative emotions.

H1a: Government parties should exhibit the highest level of positive emotional appeals, followed by mainstream opposition parties. Radical parties show the lowest level of positive emotional appeals.

H1b: Government parties show the lowest level of negative emotional appeals, followed by mainstream opposition parties. Radical parties show the highest level of positive emotional appeals.

The first set of hypotheses expect mainly differences in valence (positive versus negative) between party groups. However, our tools should also be able to investigate change in appeals to distinct emotions. The next set of hypotheses therefore concerns electoral campaigns. How do political parties change their communication when elections approach?

In the main text, we based our study on existing literature that shows that emotions and emotional content of messages matter for citizens' attitudes and behavior (Brader, 2005, 2006; Druckman & McDermott, 2008; G. E. Marcus et al., 2000; Utych, 2018; Valentino et al., 2009; Vasilopoulos, Marcus, & Foucault, 2018; Vasilopoulos, Marcus, Valentino, et al., 2018). Furthermore, this literature shows discrete emotions matter, because different emotions can carry diverging political consequences. If parties therefore are strategic about the usage of emotional appeals in their communication, we expect them to appeal to specific emotions that benefit them electorally. This should especially true for campaign periods where parties try to mobilize as much support as possible.

Research in political psychology has found that two emotions are particularly important for the mobilization of support in elections: enthusiasm and hope. Hope with respect to a candidate can be a crucial factor in voting preferences and can increase the consumption of campaign communication (Just et al., 2007). It is especially influential during elections as it links the individual's goals for the future with the democratic process (Kinder, 1994). Enthusiasm, on the other hand, has been found numerous times to increase political participation and the chance of turning out to vote (Brader, 2005, 2006; G. E. Marcus & Mackuen, 1993; Valentino et al., 2011). We therefore expect that parties make use of these two emotions during campaigns in order to increase electoral support in the final weeks before the election.

Turning to negative emotions, literature in political communication has found solid evidence for a so-called "electoral backlash" caused by negative campaigning (Fridkin & Kenney, 2011; Gerstlé & Nai, 2019; Jasperson & Fan, 2002; Lau & Pomper, 2004). The idea of a backlash is that citizens start to perceive politicians in a negative light if they make too much use of negative advertising or if they attack their opponents in an unfair and uncivil manner. Backlash would entail that citizens then become more likely to vote against that candidate and thereby punish negative campaigning. Yet, a large strand of literature shows that the effects are dependent on a number of factors.

Research shows that backlash effects depend significantly on the personalities of the recipients (Nai & Maier, 2020). For instance, individuals high in conflict avoidance show more negative evaluations of the sponsor compared to recipients low in conflict avoidance. Another factor influencing backlash effects is the partisanship of individuals (Haselmayer et al., 2020; Muddiman, 2017), whereas the negative messages of politicians from one's own party are perceived as less negative or uncivil compared to negative messages from the opposite camp. Backlash effects can also be influenced by the strength of civility/incivility (Mutz & Reeves, 2005), by the incumbency status of the sponsor (Fridkin & Kenney, 2011), and the focus of the attacks (trait or person-based) (Carraro et al., 2010). While all these studies hint towards negative consequences of negative campaigning for the sponsor, studies also show that strong, uncivil personal attacks can result in positive consequences in terms of electoral support (Gerstlé & Nai, 2019) or political engagement of the public (Brooks & Geer, 2007).

We, nevertheless, expect that parties decrease negative emotional appeals during the campaign period. We base this expectation on studies that show that negative campaigning is a strategy more often used by 'rank-and-file politicians' and party backbenchers (Dolezal et al., 2017; Haselmayer et al., 2019). Party leaders or high party officials, on the other hand, use negative campaigning less which can be explained by several reasons. Firstly, high ranking officials are the most visible party representatives and can, in case of a public backlash, cause more damage to the party than party backbenchers (Dolezal et al., 2017). Parties therefore often implement a 'division of labor' where lower-ranked politicians lead the attacks against opponents. Furthermore, assuming that parties aim for high levels of media attention, higher party officials already have more newsworthiness (due to their higher position) compared to rank-and-file politicians and hence do not need to create more newsworthiness through negative campaigning (Haselmayer et al., 2019).

Since we are analyzing press releases (see discussion of press releases below) from official party websites, we expect that party strategists within these parties try to minimize negativity during election periods in these text documents. We therefore expect lower levels of negative emotions during election campaigns in comparison to routine times. However, it is important to note that these expectations are specifically tailored towards the data and communication channel at hand. If researchers would be analyzing, for instance, social media data from party backbenches, the expectations might look differently.

H2a: Parties should during election campaigns increase appeals to discrete emotions that are mobilizing and that increase electoral support (enthusiasm and hope).

H2b: Parties should during election campaigns decrease appeals to negative emotions in order to avoid electoral backlash.

M.2 Data & Analysis

To collect press releases, we scraped the official websites of the political parties and the websites of the parties' parliamentary groups. Press releases constitute an ideal format to investigate the strategic usage of emotional language by political parties. Firstly, unlike party manifestos, press releases are released often on a daily basis and not only in election years. This enables us to compare routine and campaign periods. Moreover, there are no formal constraints imposed upon a press release's content or style (Grimmer, 2010). Parties can therefore use press releases to attack opponents or mobilize their supporters using highly emotional language. Furthermore, unlike parliamentary speeches, press releases are not constrained by the legislative agendas (Klüver & Sagarzazu, 2016). Parties can independently choose the topics they want to communicate to the voters. However, first and foremost, the purpose of press releases is to attract media attention. In general, journalists routinely rely on press releases as they provide free and easily usable information for the media (Shoemaker & Reese, 2013). For politicians, news media represent important communication channels, as they still represent one of the most important sources of information for voters during elections (Meyer et al., 2020; Strömbäck & Van Aelst, 2013). And studies show that they politicians benefit electorally from increased media attention (Gerstlé & Nai, 2019; Schaffner, 2006).

Politicians should therefore have an interest in increasing the number of messages that are being covered by news media, which might impose certain rules to the successful usage of press releases. Yet, factors that influence which messages make it into the news are multifaceted and research points into different directions. Studies found, for instance, that the status level of the politician (Flowers et al., 2003) as well as the content or topic of press releases (Meyer et al.,

2020) matter for media attention. Moreover, a study from Austria shows a partisan bias: newspapers are more likely to cover press releases from parties their readers favor (Haselmayer et al., 2017). In terms of negativity/positivity, some studies show how negative campaigning can increase media attention (e.g. Haselmayer et al., 2019). Yet, a study on campaigning worldwide reveals that more positive campaigning and appeals to specific positive emotions (enthusiasm) increase media coverage (Gerstlé & Nai, 2019), even though "nasty" personal attacks can still arouse media attention.

To sum up, we believe that press releases are an ideal medium for studying strategic emotional communication. As the research outlined above shows, politicians and parties should be strategic about the framing of their messages since the content and the tone of press releases can matter for media attention and electoral results.

As mentioned above, we collected all press releases from six German parties between January 1, 2016 and December 31, 2018. We consider campaign communication as press releases that have been released within three months prior to the elections on September 24, 2017. After data collection we split the press releases into sentences since the Electra model has been trained on the sentence level.

M.3 Findings

Figure M.1 and M.2 graphically display the results concerning the first set of hypotheses. We expected radical parties to show the highest level of negative emotions in their press releases. Conversely, we expected government parties to show the highest level of positive emotional language in their press release. Mainstream opposition parties are expected to be in the middle ground.

Figure M.1 shows the differences between party groups for negative emotions. The plot provides a first indication that our hypotheses are correct. As shown, radical parties on the fringes of the political spectrum in Germany show significantly higher levels of negative emotional appeals in their press releases compared to mainstream parties. Among negative emotions, anger is the emotion that radical parties appeal the most to, followed by fear. This is in line with literature showing the importance of these two emotions for populist radical parties (Rico et al., 2017; Vasilopoulos, Marcus, Valentino, et al., 2018). As expected, mainstream parties show significantly lower levels of negative emotional appeals. Government parties show the lowest level of anger and fear, whereas opposition parties lie in between radicals and incumbents for these emotions. However, different than expected opposition and government parties show similar levels of disgust and sadness. This finding might be explained due to the fact that mainstream parties routinely switch from government to opposition (De Vries & Hobolt, 2020) and therefore refrain from framing the political system overly negative (for example by using harsh disgust related language).

Figure M.2 shows the same results for positive emotions. As can be seen, our hypotheses are fully confirmed. Government parties show by far the highest level of positive emotions, with the highest salience of hope related language. Government parties are followed by mainstream opposition parties, that use significantly less positive emotional appeals in their press releases. Lastly, radical parties show the lowest level of positive emotions in their political communication.

Drawing on the effects of discrete emotions, we expected in our second set of hypotheses that political parties appeal to specific positive emotions in the campaign communication. On the other hand, we expected political parties to decrease negative emotional appeals in their communication in order to avoid a electoral backlash. Figure M.3 and M.4 present the results of this analysis. As can be seen, our hypotheses are largely met. Political parties decrease negative emotional appeals in the press releases during a campaign period (with the exception of anger which does not reach statistical significance). On the other hand, as expected, parties appeal strategically to enthusiasm and hope, and not to other positive emotions, in order to increase their electoral support.

Figure M. 1: Emotions (ELECTRA) by party group (negative emotions)

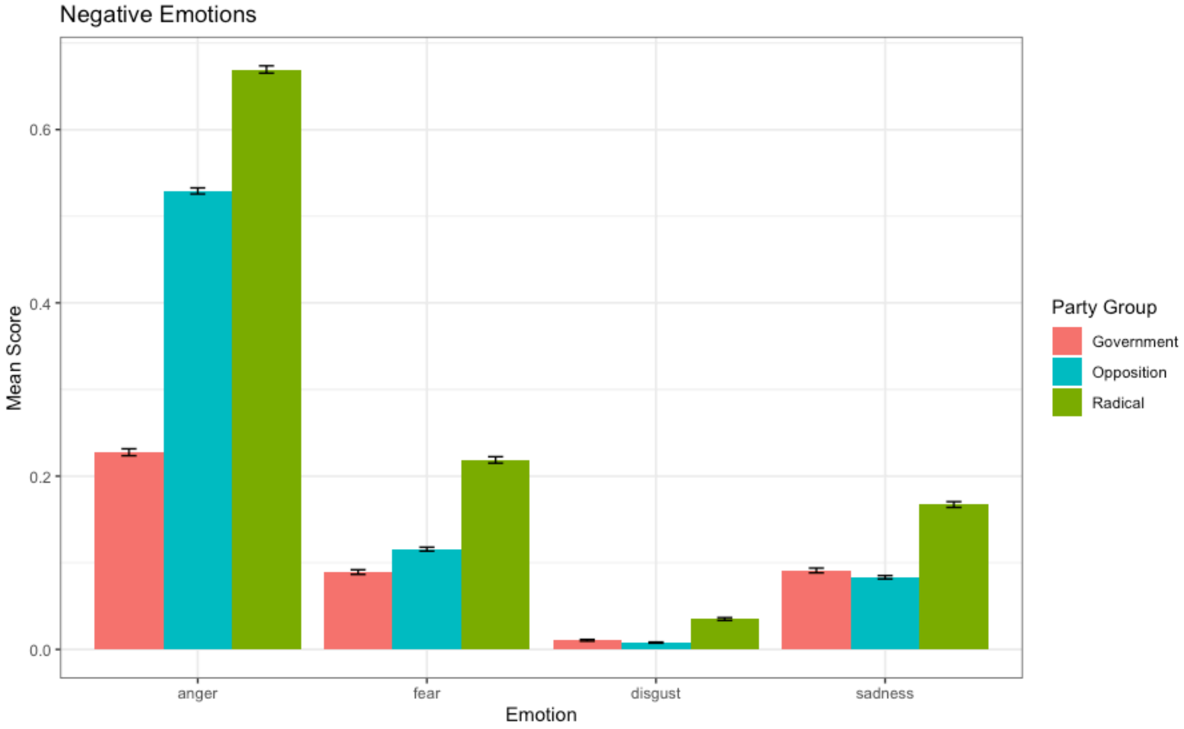


Figure M. 2: Emotions (ELECTRA) by party group (positive emotions)

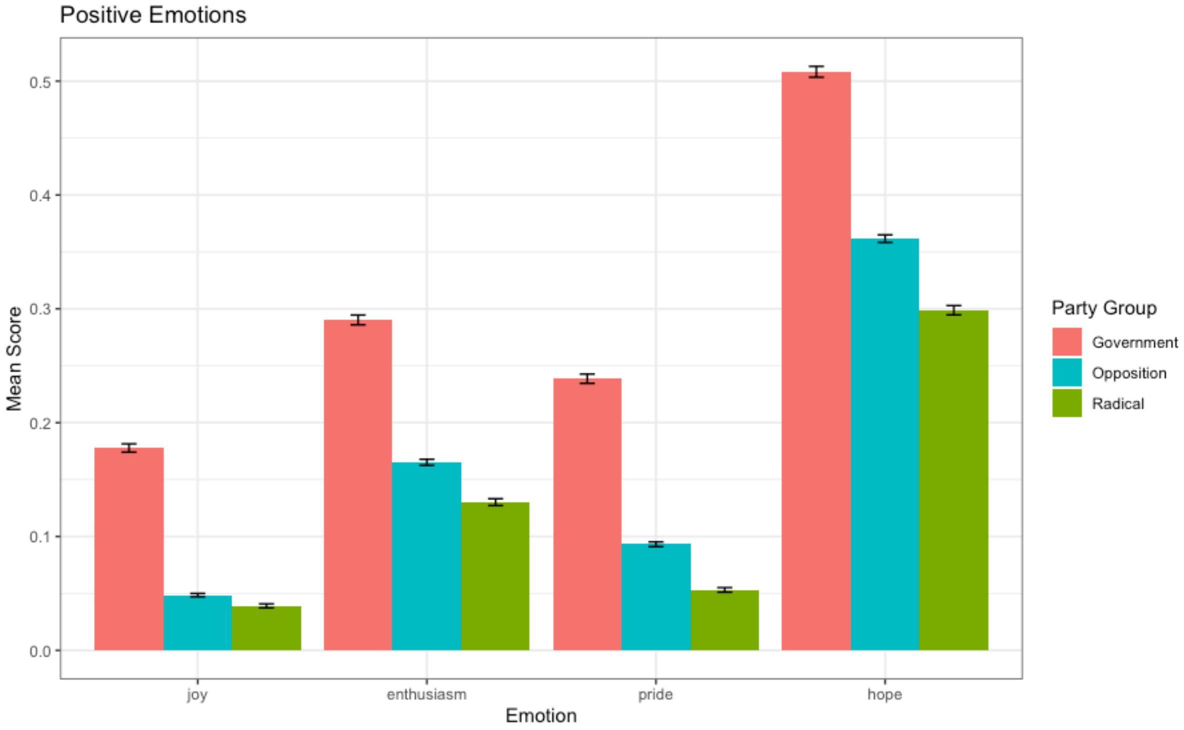


Figure M. 3: Effect of campaign period on emotional appeals (negative emotions)

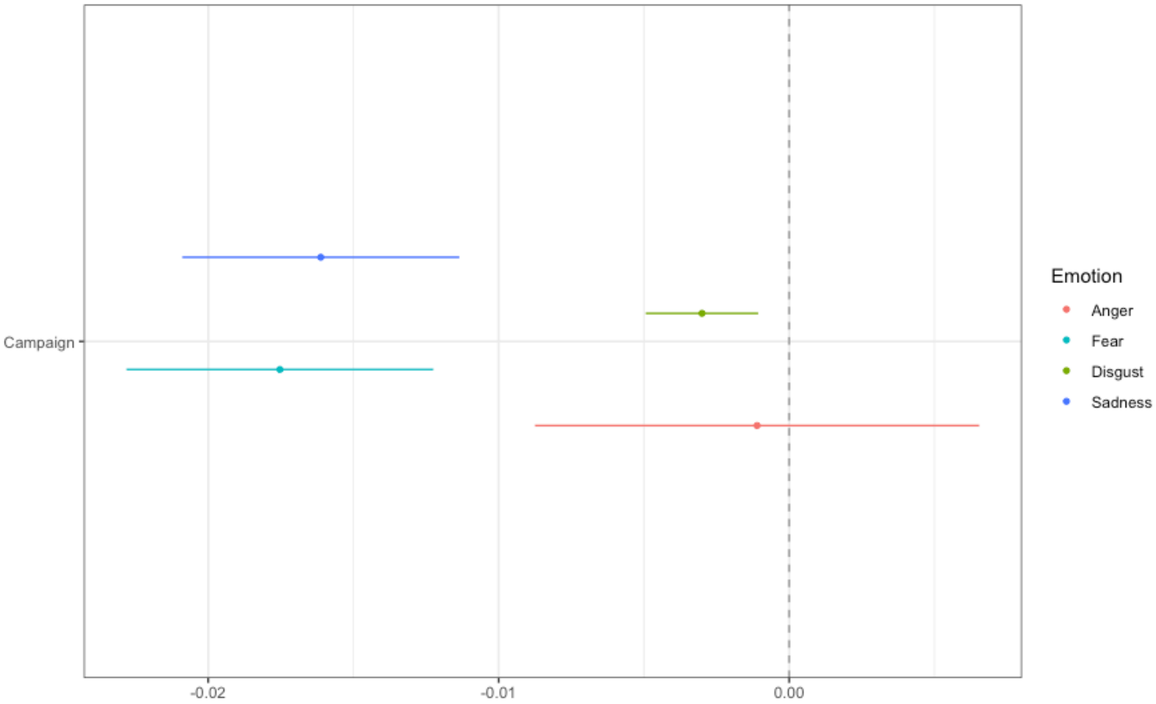
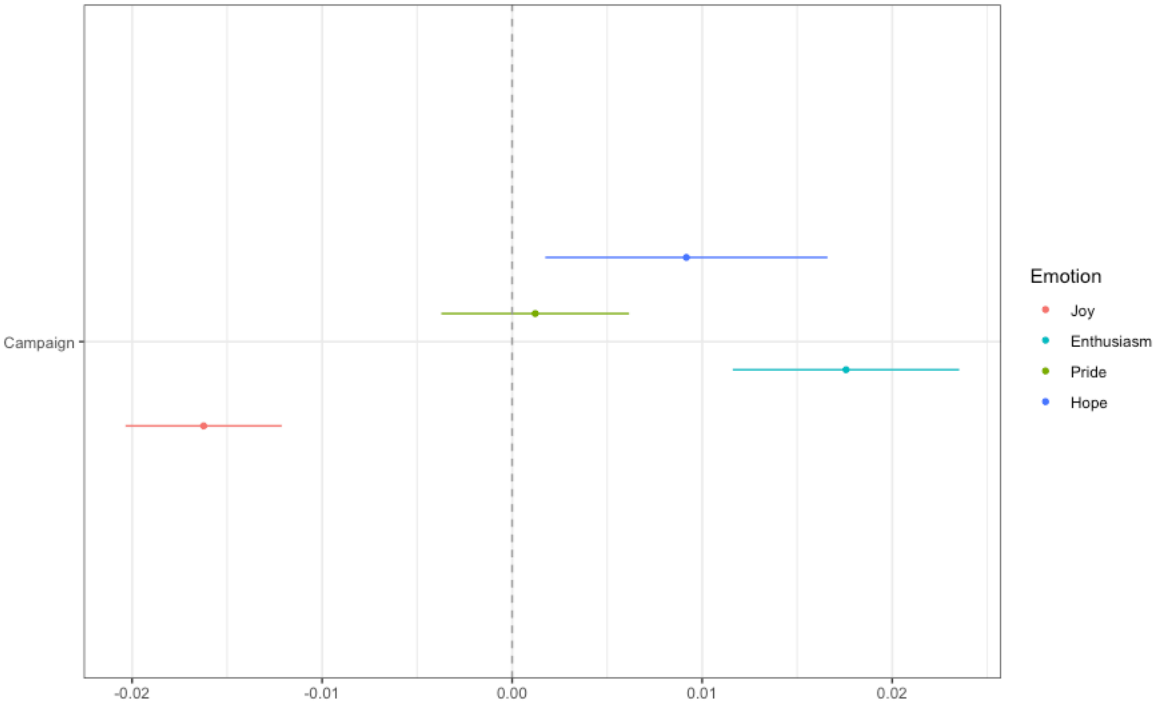


Figure M. 4: Effect of campaign period on emotional appeals (positive emotions)



References Online Appendix

Aarøe, L., Petersen, M. B., & Arceneaux, K. (2017). The Behavioral Immune System Shapes Political Intuitions: Why and How Individual Differences in Disgust Sensitivity Underlie Opposition to Immigration. *American Political Science Review*, *111*(2), 277–294. https://doi.org/10.1017/S0003055416000770

Alhuzali, H., & Ananiadou, S. (2021). SpanEmo: Casting Multi-label Emotion Classification as Span-prediction. *ArXiv Preprint ArXiv:2101.10038*.

Al-Rfou', R., Perozzi, B., & Skiena, S. (2013). Polyglot: Distributed Word Representations for Multilingual NLP. *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, 183–192.

Andreevskaia, A., & Bergler, S. (2006, April). Mining WordNet for a Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses. *11th Conference of the European Chapter of the Association for Computational Linguistics*. EACL 2006, Trento, Italy. https://www.aclweb.org/anthology/E06-1027

Arzheimer, K., & Berning, C. C. (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

Banks, A. J., & Valentino, N. A. (2012). Emotional Substrates of White Racial Attitudes. *American Journal of Political Science*, *56*(2), 286–297. https://doi.org/10.1111/j.1540-5907.2011.00561.x

Benoit, K., Conway, D., Lauderdale, B. E., Laver, M., & Mikhaylov, S. (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

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2014). Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys. *American Journal of Political Science*, *58*(3), 739–753. https://doi.org/10.1111/ajps.12081

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *ArXiv:1607.04606 [Cs]*. http://arxiv.org/abs/1607.04606

Brader, T. (2005). Striking a Responsive Chord: How Political Ads Motivate and Persuade Voters by Appealing to Emotions. *American Journal of Political Science*, *49*(2), 388–405. https://doi.org/10.1111/j.0092-5853.2005.00130.x

Brader, T. (2006). *Campaigning for hearts and minds: How emotional appeals in political ads work*. University of Chicago Press.

Brader, T., & Marcus, G. E. (2013). Emotion and Political Psychology. *The Oxford Handbook of Political Psychology*. https://doi.org/10.1093/oxfordhb/9780199760107.013.0006

Brooks, D. J., & Geer, J. G. (2007). Beyond Negativity: The Effects of Incivility on the Electorate. *American Journal of Political Science*, *51*(1), 1–16. https://doi.org/10.1111/j.1540-5907.2007.00233.x

Carraro, L., Gawronski, B., & Castelli, L. (2010). Losing on all fronts: The effects of negative versus positive person-based campaigns on implicit and explicit evaluations of political candidates. *British Journal of Social Psychology*, *49*(3), 453–470. https://doi.org/10.1348/014466609X468042

Crabtree, C., Golder, M., Gschwend, T., & Indriđason, I. H. (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

De Vries, C. E., & Hobolt, S. B. (2020). *Political entrepreneurs: The rise of challenger parties in Europe*. Princeton University Press.

De Zavala, A. G., Cichocka, A., Eidelson, R., & Jayawickreme, N. (2009). Collective narcissism and its social consequences. *Journal of Personality and Social Psychology*, *97*(6), 1074.

Desteno, D., Wegener, D. T., Petty, R. E., Rucker, D. D., & Braverman, J. (2004). *Discrete Emotions and Persuasion: The Role of Emotion-induced Expectancies*.

Dolezal, M., Ennser-Jedenastik, L., & Müller, W. C. (2017). Who will attack the competitors? How political parties resolve strategic and collective action dilemmas in negative campaigning. *Party Politics*, *23*(6), 666–679. https://doi.org/10.1177/1354068815619832

Druckman, J. N., & McDermott, R. (2008). Emotion and the framing of risky choice. *Political Behavior*, *30*(3), 297–321.

Faircrowd.com. (2017). *Ombuds Office for German crowdsourcing platforms established -Fair Crowd Work*. http://faircrowd.work/2017/11/08/ombudsstelle-fuer-crowdworking-plattformen-vereinbart/

Flowers, J. F., Haynes, A. A., & Crespin, M. H. (2003). The Media, the Campaign, and the Message. *American Journal of Political Science*, *47*(2), 259–273. https://doi.org/10.1111/1540-5907.00018

Fort, K., Adda, G., & Cohen, K. B. (2011). Amazon Mechanical Turk: Gold Mine or Coal Mine? *Computational Linguistics*, *37*(2), 413–420. https://doi.org/10.1162/COLI_a_00057

Fridkin, K. L., & Kenney, P. (2011). Variability in Citizens' Reactions to Different Types of Negative Campaigns. *American Journal of Political Science*, *55*(2), 307–325. https://doi.org/10.1111/j.1540-5907.2010.00494.x

Gadarian, S. K., & Albertson, B. (2014). Anxiety, Immigration, and the Search for Information: Anxiety, Immigration, and the Search for Information. *Political Psychology*, *35*(2), 133–164. https://doi.org/10.1111/pops.12034

Gebert, M. (2017). *Crowdsourcing: Code of Conduct*. http://crowdsourcing-code.com/

Gerstlé, J., & Nai, A. (2019). Negativity, emotionality and populist rhetoric in election campaigns worldwide, and their effects on media attention and electoral success. *European Journal of Communication*, *34*(4), 410–444. https://doi.org/10.1177/0267323119861875

Grimmer, J. (2010). A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases. *Political Analysis*, *18*(1), 1–35.

Grzesiak-Feldman, M. (2013). The Effect of High-Anxiety Situations on Conspiracy Thinking. *Current Psychology*, *32*(1), 100–118. https://doi.org/10.1007/s12144-013-9165-6

Hameleers, M., Bos, L., & de Vreese, C. H. (2016). “They Did It”: The Effects of Emotionalized Blame Attribution in Populist Communication. *Communication Research*, *44*(6), 870–900. https://doi.org/10.1177/0093650216644026

Hara, K., Adams, A., Milland, K., Savage, S., Callison-Burch, C., & Bigham, J. (2017). A Data-Driven Analysis of Workers’ Earnings on Amazon Mechanical Turk. *ArXiv:1712.05796 [Cs]*. http://arxiv.org/abs/1712.05796

Haselmayer, M., Hirsch, L., & Jenny, M. (2020). Love is blind. Partisanship and perception of negative campaign messages in a multiparty system. *Political Research Exchange*, *2*(1), 1806002. https://doi.org/10.1080/2474736X.2020.1806002

Haselmayer, M., Meyer, T. M., & Wagner, M. (2019). Fighting for attention: Media coverage of negative campaign messages. *Party Politics*, *25*(3), 412–423. https://doi.org/10.1177/1354068817724174

Haselmayer, M., Wagner, M., & Meyer, T. M. (2017). Partisan Bias in Message Selection: Media Gatekeeping of Party Press Releases. *Political Communication*, *34*(3), 367–384. https://doi.org/10.1080/10584609.2016.1265619

Healy, A. J., Malhotra, N., & Mo, C. H. (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

Hobolt, S. B., & de Vries, C. E. (2015). Issue Entrepreneurship and Multiparty Competition. *Comparative Political Studies*, *48*(9), 1159–1185. https://doi.org/10.1177/0010414015575030

Huddy, L., Smirnov, O., Snider, K. L. G., & Perliger, A. (2021). Anger, Anxiety, and Selective Exposure to Terrorist Violence. *Journal of Conflict Resolution*, 00220027211014937. https://doi.org/10.1177/00220027211014937

Imai, K., & Khanna, K. (2016). Improving Ecological Inference by Predicting Individual Ethnicity from Voter Registration Records. *Political Analysis*, *24*(2), 263–272.

Inbar, Y., Pizarro, D. A., & Bloom, P. (2009). Conservatives are more easily disgusted than liberals. *Cognition and Emotion*, *23*(4), 714–725. https://doi.org/10.1080/02699930802110007

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

Jasperson, A. E., & Fan, D. P. (2002). An aggregate examination of the backlash effect in political advertising: The case of the 1996 US Senate race in Minnesota. *Journal of Advertising*, *31*(1), 1–12.

Just, M. R., Crigler, A. N., & Belt, T. L. (2007). Don't give up hope: Emotions, candidate appraisals, and votes. In W. R. Neuman, G. E. Marcus, A. N. Crigler, & M. MacKuen (Eds.), *The affect effect: Dynamics of emotion in political thinking and behavior*. The University of Chicago Press.

Kinder, D. R. (1994). Reason and Emotion in American Political Life. In *Beliefs, Reasoning, and Decision Making*. Psychology Press.

Klüver, H., & Sagarzazu, I. (2016). Setting the Agenda or Responding to Voters? Political Parties, Voters and Issue Attention. *West European Politics*, *39*(2), 380–398. https://doi.org/10.1080/01402382.2015.1101295

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

Lau, R. R., & Pomper, G. M. (2004). *Negative campaigning: An analysis of US Senate elections*. Rowman & Littlefield.

Lazarus, R. S. (1991). *Emotion and adaptation*. Oxford University Press.

Lerner, J. S., Gonzalez, R. M., Small, D. A., & Fischhoff, B. (2003). Effects of Fear and Anger on Perceived Risks of Terrorism: A National Field Experiment. *Psychological Science*, *14*(2), 144–150. https://doi.org/10.1111/1467-9280.01433

Lerner, J. S., & Keltner, D. (2000). Beyond valence: Toward a model of emotion-specific influences on judgement and choice. *Cognition & Emotion*, *14*(4), 473–493.

Lerner, J. S., & Keltner, D. (2001). Fear, anger, and risk. *Journal of Personality and Social Psychology*, *81*(1), 146.

Lind, F., Gruber, M., & Boomgaarden, H. G. (2017). Content Analysis by the Crowd: Assessing the Usability of Crowdsourcing for Coding Latent Constructs. *Communication Methods and Measures*, *11*(3), 191–209. https://doi.org/10.1080/19312458.2017.1317338

MacKuen, M., Wolak, J., Keele, L., & Marcus, G. E. (2010). Civic Engagements: Resolute Partisanship or Reflective Deliberation. *American Journal of Political Science*, *54*(2), 440–458. https://doi.org/10.1111/j.1540-5907.2010.00440.x

Marcus, G. E., & Mackuen, M. B. (1993). Anxiety, Enthusiasm, and the Vote: The Emotional Underpinnings of Learning and Involvement During Presidential Campaigns. *The American Political Science Review*, *87*(3), 672–685. JSTOR. https://doi.org/10.2307/2938743

Marcus, G. E., MacKuen, M., Wolak, J., & Keele, L. (2006). The measure and mismeasure of emotion. In *Feeling politics* (pp. 31–45). Springer.

Marcus, G. E., Neuman, W. R., & MacKuen, M. (2000). *Affective intelligence and political judgment*. University of Chicago Press.

Marcus, G., Neuman, W. R., & MacKuen, M. (2017). Measuring Emotional Response: Comparing Alternative Approaches to Measurement. *Journal of Political Science Research and Methods*, *5*, 733–754.

Matsumoto, D., Frank, M. G., & Hwang, H. C. (2015). The Role of Intergroup Emotions in Political Violence. *Current Directions in Psychological Science*, *24*(5), 369–373. https://doi.org/10.1177/0963721415595023

Meyer, T. M., Haselmayer, M., & Wagner, M. (2020). Who Gets into the Papers? Party Campaign Messages and the Media. *British Journal of Political Science*, *50*(1), 281–302. https://doi.org/10.1017/S0007123417000400

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2017). Advances in Pre-Training Distributed Word Representations. *ArXiv:1712.09405 [Cs]*. http://arxiv.org/abs/1712.09405

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *ArXiv Preprint ArXiv:1310.4546*.

Mnih, A., & Kavukcuoglu, K. (2013). Learning word embeddings efficiently with noise-contrastive estimation. *Advances in Neural Information Processing Systems*, *26*, 2265–2273.

Muchlinski, D., Siroky, D., He, J., & Kocher, M. (2016). Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data. *Political Analysis*, *24*(1), 87–103.

Mudde, C. (2004). The populist zeitgeist. *Government and Opposition*, *39*(4), 541–563.

Muddiman, A. (2017). *Personal and Public Levels of Political Incivility*. 21.

Mutz, D. C., & Reeves, B. (2005). The New Videomalaise: Effects of Televised Incivility on Political Trust. *American Political Science Review*, *99*(1), 1–15. https://doi.org/10.1017/S0003055405051452

Nai, A., & Maier, J. (2020). Is Negative Campaigning a Matter of Taste? Political Attacks, Incivility, and the Moderating Role of Individual Differences. *American Politics Research*, 1532673X20965548. https://doi.org/10.1177/1532673X20965548

Newman, A. (2019, November 15). I Found Work on an Amazon Website. I Made 97 Cents an Hour. *The New York Times*. https://www.nytimes.com/interactive/2019/11/15/nyregion/amazon-mechanical-turk.html, https://www.nytimes.com/interactive/2019/11/15/nyregion/amazon-mechanical-turk.html

O'Connor, S. (2020, September 15). Do not let homeworking become digital piecework for the poor. *Financial Times*. https://www.ft.com/content/ab83270c-253a-4d7a-9ef8-a360d2e04aab

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological Aspects of Natural Language Use: Our Words, Our Selves. *Annual Review of Psychology*, *54*(1), 547–577. https://doi.org/10.1146/annurev.psych.54.101601.145041

Proksch, S.-O., & Slapin, J. B. (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*, *0*(0), 1–25. https://doi.org/10.1080/19331681.2018.1485608

Rico, G., Guinjoan, M., & Anduiza, E. (2017). The Emotional Underpinnings of Populism: How Anger and Fear Affect Populist Attitudes. *Swiss Political Science Review*, *23*(4), 444–461. https://doi.org/10.1111/spsr.12261

Rooduijn, M., Van Kessel, S., Froio, C., Pirro, A., De Lange, S., Halikiopoulou, D., Lewis, P., Mudde, C., & Taggart, P. (2019). *The PopuList: An Overview of Populist, Far Right, Far Left and Eurosceptic Parties in Europe.* www.popu-list.org

Ross, J., Zaldivar, A., Irani, L., & Tomlinson, B. (2010). *Who are the Turkers? Worker Demographics in Amazon Mechanical Turk*. 5.

Salmela, M., & von Scheve, C. (2018). Emotional Dynamics of Right- and Left-wing Political Populism. *Humanity & Society*, *42*(4), 434–454. https://doi.org/10.1177/0160597618802521

Salmond, R. (2014). Parliamentary Question Times: How Legislative Accountability Mechanisms Affect Mass Political Engagement. *The Journal of Legislative Studies*, *20*(3), 321–341. https://doi.org/10.1080/13572334.2014.895121

Schaffner, B. F. (2006). Local news coverage and the incumbency advantage in the US House. *Legislative Studies Quarterly*, *31*(4), 491–511.

Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2008). Disgust as embodied moral judgment. *Personality and Social Psychology Bulletin*, *34*(8), 1096–1109.

Shmueli, B., Fell, J., Ray, S., & Ku, L.-W. (2021). Beyond Fair Pay: Ethical Implications of NLP Crowdsourcing. *ArXiv:2104.10097 [Cs]*. http://arxiv.org/abs/2104.10097

Shoemaker, P. J., & Reese, S. D. (2013). *Mediating the message in the 21st century: A media sociology perspective*. Routledge.

Small, D. A., & Lerner, J. S. (2008). Emotional Policy: Personal Sadness and Anger Shape Judgments about a Welfare Case. *Political Psychology*, *29*(2), 149–168. https://doi.org/10.1111/j.1467-9221.2008.00621.x

Smith, C. A., & Ellsworth, P. C. (1985). Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology*, *48*(4), 813.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, *15*(56), 1929–1958.

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/

Stewart, B. M., & Zhukov, Y. M. (2009). Use of force and civil–military relations in Russia: An automated content analysis. *Small Wars & Insurgencies*, *20*(2), 319–343.

Stier, S., Posch, L., Bleier, A., & Strohmaier, M. (2017). When populists become popular: Comparing Facebook use by the right-wing movement Pegida and German political parties. *Information, Communication & Society*, *20*(9), 1365–1388.

Strömbäck, J., & Van Aelst, P. (2013). Why political parties adapt to the media: Exploring the fourth dimension of mediatization. *International Communication Gazette*, *75*(4), 341–358. https://doi.org/10.1177/1748048513482266

Subasic, P., & Huettner, A. (2001). Affect analysis of text using fuzzy semantic typing. *IEEE Transactions on Fuzzy Systems*, *9*(4), 483–496. https://doi.org/10.1109/91.940962

Tiedens, L. Z., & Linton, S. (2001). *Judgment Under Emotional Certainty and Uncertainty: The Effects of Specific Emotions on Information Processing*. 16.

Tresch, A. (2009). Politicians in the Media: Determinants of Legislators' Presence and Prominence in Swiss Newspapers. *The International Journal of Press/Politics*, *14*(1), 67–90. https://doi.org/10.1177/1940161208323266

Turner, J. H. (2007). Self, emotions, and extreme violence: Extending symbolic interactionist theorizing. *Symbolic Interaction*, *30*(4), 501–530.

Utych, S. M. (2018). Negative Affective Language in Politics. *American Politics Research*, *46*(1), 77–102. https://doi.org/10.1177/1532673X17693830

Valentino, Brader, T., Groenendyk, E. W., Gregorowicz, K., & Hutchings, V. L. (2011). Election night's alright for fighting: The role of emotions in political participation. *The Journal of Politics*, *73*(1), 156–170.

Valentino, N. A., Gregorowicz, K., & Groenendyk, E. W. (2009). Efficacy, Emotions and the Habit of Participation. *Political Behavior*, *31*(3), 307–330.

Vasilopoulos, P., Marcus, G. E., & Foucault, M. (2018). Emotional Responses to the Charlie Hebdo Attacks: Addressing the Authoritarianism Puzzle. *Political Psychology*, *39*(3), 557–575. https://doi.org/10.1111/pops.12439

Vasilopoulos, P., Marcus, G. E., Valentino, N. A., & Foucault, M. (2018). Fear, Anger, and Voting for the Far Right: Evidence From the November 13, 2015 Paris Terror Attacks. *Political Psychology*, *0*(0). https://doi.org/10.1111/pops.12513

Vasilopoulou, S., & Wagner, M. (2017). Fear, anger and enthusiasm about the European Union: Effects of emotional reactions on public preferences towards European integration. *European Union Politics*, *18*(3), 382–405. https://doi.org/10.1177/1465116517698048

Wang, Chen, L., Thirunarayan, K., & Sheth, A. P. (2012). Harnessing Twitter "Big Data" for Automatic Emotion Identification. *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, 587–592. https://doi.org/10.1109/SocialCom-PASSAT.2012.119

Widmann, T. (2021). How Emotional Are Populists Really? Factors Explaining Emotional Appeals in the Communication of Political Parties. *Political Psychology*, *42*(1), 163–181. https://doi.org/10.1111/pops.12693

Williamson, V. (2016). On the ethics of crowdsourced research. *PS: Political Science & Politics*, *49*(1), 77–81.

Wolf, T., Chaumond, J., Debut, L., Sanh, V., Delangue, C., Moi, A., Cistac, P., Funtowicz, M., Davison, J., & Shleifer, S. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45.

Young, L., & Soroka, S. (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