PA

Sentiment is Not Stance: Target-Aware Opinion Classification for Political Text Analysis

Samuel E. Bestvater and Burt L. Monroe

Department of Political Science, The Pennsylvania State University, University Park, PA, USA. E-mail: seb654@psu.edu, burtmonroe@psu.edu

Abstract

Sentiment analysis techniques have a long history in natural language processing and have become a standard tool in the analysis of political texts, promising a conceptually straightforward automated method of extracting meaning from textual data by scoring documents on a scale from positive to negative. However, while these kinds of sentiment scores can capture the overall tone of a document, the underlying concept of interest for political analysis is often actually the document's stance with respect to a given target—how positively or negatively it frames a specific idea, individual, or group—as this reflects the author's underlying political attitudes. In this paper, we question the validity of approximating author stance through sentiment scoring in the analysis of political texts, and advocate for greater attention to be paid to the conceptual distinction between a document's sentiment and its stance. Using examples from open-ended survey responses and from political discussions on social media, we demonstrate that in many political text analysis applications, sentiment and stance do not necessarily align, and therefore sentiment analysis methods fail to reliably capture ground-truth document stance, amplifying noise in the data and leading to faulty conclusions.

Keywords: text-as-data, sentiment analysis, political stance, machine learning

1 Introduction

As text-as-data methods have become increasingly popular in the social sciences, lexicon-based sentiment analysis techniques have emerged as a popular off-the-shelf tool for the automated extraction of information from political texts. The approach is both conceptually and computationally simple: many words have a recognizable positive or negative valence which can be recorded in a dictionary of term-valence pairs, then applied to any given document to produce an aggregate measure of its overall tone or polarity. This is a useful quantity to measure for a variety of subjects in political research, where we might care about the negativity of campaign ads (Hopp and Vargo 2017) or political news coverage (Soroka, Young, and Balmas 2015; Young and Soroka 2012), for example. However, for many other applications in political analysis, we are less interested in a document's overall tone and more concerned with what it says *about* something in particular. We want to be able to use the text to measure the author's underlying political attitudes about specific ideas, individuals, or groups. And although sentiment analysis is frequently applied to these types of tasks, it is not necessarily well-suited for them, and this practice can result in issues of validity that have caused many political science practitioners to view sentiment analysis with a certain degree of skepticism (e.g., González-Bailón and Paltoglou 2015; Klašnja et al. 2015). In this article, we address this skepticism through an exploration of the limitations of applying sentiment analysis techniques to measure targeted political opinions from text data. Along the way, we introduce a political science audience to a growing area of research in natural language processing (NLP) that views "stance detection" tasks as distinct from general sentiment analysis, and expand on this research to explicitly consider the importance of accurately conceptualizing and measuring stance for use in downstream analyses of political phenomena. Through a better understanding of this conceptual distinction between sentiment and stance, as

Political Analysis (2023)
vol. 31: 235–256
DOI: 10.1017/pan.2022.10

Published
22 April 2022

Corresponding author
Samuel E. Bestvater

Edited by
Jeff Gill

well as the tools appropriate to each, researchers can improve the validity of opinion measures extracted from political texts and have more confidence in the inferences drawn from such measures.

The central argument in this article rests on a fairly nuanced conceptual point, which is that a document's *sentiment* is not equivalent to its *stance*, so it is useful to begin by laying out some definitions. Automated sentiment analysis is a core task in NLP, and is concerned with classifying or scaling a document according to its general polarity on a positive–negative scale (Küçük and Can 2020; Pang, Lee, and Vaithyanathan 2002). By contrast, stance is an affective or attitudinal position expressed toward a given target: a document's negative–positive polarity not in general, but as defined relative to a specific concept or entity of interest (Mohammad et al. 2016). Researchers in NLP have come to view the problem of automated stance detection as distinct from general sentiment analysis (Küçük and Can 2020), and have recently devoted greater attention to the problem of identifying the polarity of targeted opinions (e.g., Abercrombie et al. 2019). As a linguistic concept, however, stance itself dates back much further. Biber and Finegan (1988) define stance as the expression of a speaker's standpoint and judgment toward a given proposition, while Du Bois (2007) describes it as "a public act by a social actor, achieved dialogically through overt communicative means, of simultaneously evaluating objects, positioning subjects (self and others), and aligning with other subjects, with respect to any salient dimension of the sociocultural field."

In these definitions, stance is distinct from sentiment because it is targeted—it pertains to something specific rather than simply reflecting the overall tone of the language used in the document. However, even if document sentiment is a general, target-agnostic concept, texts generally have topics, which prompts the question: is it reasonable to assume that the sentiment of a document reflects the stance of that document toward the primary topic of the document? In many cases, this can be perfectly reasonable. One of the canonical applications for sentiment analysis, for example, is analyzing product or movie reviews (e.g., Maas et al. 2011; Pang *et al.* 2002). The overall purpose of a review is to state and justify a stance on the subject of the review—a film was good, a product is faulty, etc. Therefore, the document's sentiment is very likely to be closely aligned to its stance. A film review saying "I absolutely loved this film" is using positive sentiment and indicating a stance in favor of the movie. By contrast, a review saying "this movie was horrible, save your money" is using negative sentiment and indicating a stance against the movie. So far so good.

However, problems can emerge when the relationship between a document's general sentiment and its stance toward a specific target is less direct, which is why stance detection has come to be viewed as a distinct task in NLP. This distinction is of particular relevance to the analysis of political discourse, as politics is often contentious and emotionally charged, so we frequently employ sentiment-laden language when we talk about it. Political attitudes also tend to be multifaceted and can almost always be expressed in a variety of ways. On the issue of gun control, someone who is pro-gun rights is likely opposed to gun restrictions. On the issue of whether or not Great Britain should remain in the European Union (EU), someone who is pro-Brexit might be considered anti-EU. In cases where the same stance can be easily expressed in either supportive or oppositional terms, sentiment and stance are not cleanly aligned and cannot be treated as conceptually equivalent. Unfortunately, despite the fact that political stances are often particularly complex and are not necessarily aligned with general sentiment, the distinction between sentiment and stance receives relatively little attention in research using political text. When we want to know what people think about a political issue, a common approach is to collect documents on that topic, conduct sentiment analysis on those documents, and then make inferences about political attitudes or predict political behaviors based on those sentiment measures. In recent years, this general approach has been extensively employed to measure public

opinion on various political figures and issues. Using variations of this basic technique, scholars have studied the popularity of candidates in elections (Murthy 2015; Rezapour et al. 2017; Wang et al. 2012), as well as popular views on politicized issues such as climate change (Dahal, Kumar, and Li 2019). Other studies commonly start by measuring opinion from text corpora through sentiment analysis techniques, then use that measure to predict some other real-world political behavior or attitude such as presidential approval (O'Connor et al. 2010) or election results (Choy et al. 2011, 2012; Jose and Chooralil 2015; Tumasjan et al. 2010).

In this article, our objective is to call attention to the conceptual difference between sentiment and stance, and to demonstrate the measurement bias that can result from using sentiment measures to operationalize stance in the analysis of political texts. We do this through three real-world examples that recreate common research scenarios where sentiment analysis techniques are improperly applied to stance identification tasks. In the section that immediately follows, we replicate and extend a recent analysis by Felmlee et al. (2020) that examined opinions contained in tweets about the 2017 Women's March. Through this exercise, we illustrate that even when sentiment and stance are correlated in a corpus, using sentiment values as a proxy for stance can lead to attenuated effects. Then, we explore two further examples–a corpus of short open-ended survey responses where respondents express their opinions on Donald Trump and a corpus of tweets where authors express their opinion on the 2018 nomination of Brett Kavanaugh to the U.S. Supreme Court—where we compare the accuracy of an array of text classifiers when applied to both sentiment and stance identification tasks.[1] Each of these examples represent political discourse around controversial, emotionally laden topics, figures, or events where the positions expressed by the authors of each text can be complex and multidimensional. We find in each example that sentiment analysis techniques produce noisier, less accurate measures that appear to attenuate the relationship between stance as expressed in political texts and ground truth measures, sometimes in extreme ways. We conclude by offering some practical advice for the use of text-as-data methods in political research, arguing that researchers should always carefully consider how closely a chosen measure captures the true quantity of interest. For many applications, training a new supervised classifier on a small training set hand-labeled for the exact quantity of interest will produce a more valid measure than relying on an existing model or dictionary that was designed to identify a related, but distinct concept.

2 Example I: Sentiment and Stance in Tweets About the 2017 Women's March

Sentiment analysis techniques are frequently employed to capture the broad contours of a conversation on a topic of interest, but issues can arise when the use of positive or negative tone in the language of the conversation is interpreted as indicating a favorable or unfavorable stance on the topic at hand. Some of these issues can be illustrated through a brief replication exercise. Shortly after the 2016 presidential election in the United States, posts on Facebook and Twitter began to appear calling for women to march on Washington DC as well as other cities and towns around America and the rest of the world to protest the political agenda of the incoming Republican administration and to advocate for a broad platform of human rights and social justice. Social media served as a key platform for sharing information about the Women's March events, and when the Marches occurred on January 21, 2017, they prompted a substantial amount of online discourse. In a recent paper, Felmlee et al. (2020) collected and analyzed a sample of 2.5 million geo-coded tweets about the Marches in order to better understand the sentiment of discourse around the movement as well as how it varied geographically.

1 Data and replication scripts for each of these examples are available at Bestvater and Monroe (2022), at https://doi.org/10.7910/DVN/MUYYG4.

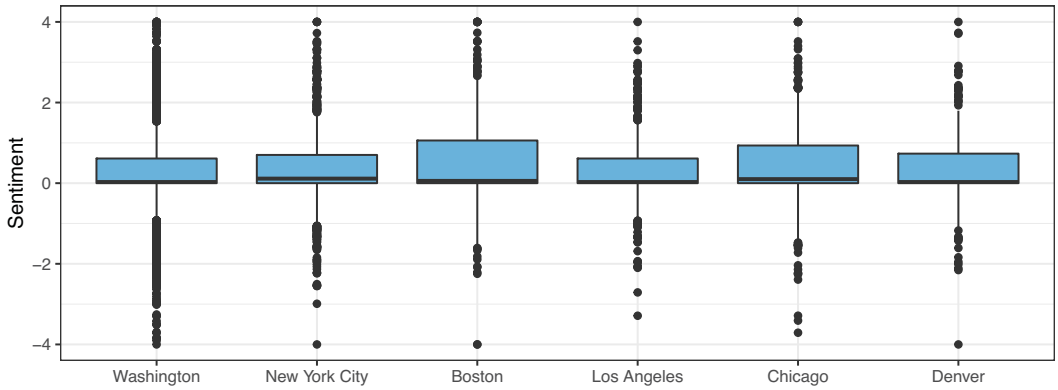

Figure 1. Geographic distribution of machine-coded sentiment in tweets about the Women's March. *Note:* Boxplots of VADER sentiment scores in geocoded tweets referencing Women's March events in the top six most-referenced cities. Replication data from Felmlee et al. (2020, Figure 7).

Felmlee et al. (2020) assign sentiment scores to the tweets they collected using VADER (Hutto and Gilbert 2014), a sentiment lexicon designed to perform well on short, informal documents such as social media posts or short answer open-ended survey responses.[2] The authors find that the tone of tweets about the Women's March was more positive on average than that of tweets on other topics, and that within the subset of tweets that mentioned Marches in specific cities, sentiment was also generally positive, with relatively little geographic variation (see Figure 1).

Felmlee et al. (2020) take these findings as evidence that the Women's March movement was an overall success, enjoying a broad geographic base of support online—a conclusion which relies on the implicit assumption that tweets about the Women's March that use positive language indicate approval, while tweets that use negative language indicate opposition. This assumption is not necessarily valid, however, as protest movements generally coalesce around grievances (see Gurr 1970) and an author's stance of approval toward the movement or its objectives can be just as easily expressed as opposition to the grievances against which the protest is directed. Likewise, opposition to a protest can be expressed as approval of the status quo as well as criticism of the movement. Fundamentally, this means that document sentiment and author stance toward a protest movement are not conceptually equivalent, and may or may not be strongly correlated.

To illustrate this point, we took a random sample of 20,000 Women's March tweets, and labeled them by hand according to whether they used generally positive or generally negative language, and whether they indicated approval or opposition for the movement.[3] Some examples from this corpus are shown in Table 1, demonstrating how tweets that indicate approval of the March can have an overall positive sentiment (talking about the movement as inspiring or rewarding,

2 Shorter documents may have fewer terms that the lexicon can identify, a problem which can be exacerbated by the use of informal language such as slang, emojis, and unconventional spelling. To overcome these common issues, VADER makes two key innovations over other sentiment lexicons. First, it uses a crowdsourced vocabulary of over 7,500 terms, which include formal English terms as well as common slang and abbreviations such as "LOL" or "OMG" as well as also emoticons and unicode emoji characters. Sentiment values for these terms were scored on a scale of −4 to 4 by 10 crowdworkers each, in order to avoid introducing bias from particular coders and to get closer to the "ground truth" of how each term is used in common language. Second, VADER does not apply this dictionary using a naive "bag of words" assumption, but follows a series of algorithmic rules. Exclamation marks serve as a multiplier, increasing the magnitude of a sentiment expressed in a sentence or document without changing its orientation. Likewise, if a sentiment-laden term appears in all caps, its sentiment score is increased as well. Intensifying adjectives and adverbs also increase the sentiment scores of the terms they modify, so that "extremely good" is scored as more positive than simply "good." Finally, VADER evaluates terms in local-window contexts of three words in order to capture negations and flip sentiment polarity accordingly, and it also reweights sentences that use tone-shifting conjunctions like "but" or "however" so that the overall sentiment reflects the shift.

3 To avoid issues of overfitting that can arise from training supervised classifiers on observations that appear in the test set, this validation sample of 20,000 tweets is not drawn from the original sample of 2.5 million tweets used in Felmlee et al. (2020). Instead, we used the Twitter API to collect a new sample of Women's March tweets from the same timeframe, filtered to exclude tweets that appear in the Felmlee et al. (2020) corpus.

Table 1. Human-labeled sentiment and stance in tweets about the Women's March.

	Positive sentiment	Negative sentiment
Approving stance	• Being able to take part in such an important movement has been so rewarding and I just hope that we are heard. #WomensMarch • Much respect to all the POWERFUL women standing and marching for their values today. #WomensMarch	• Congress—do you hear us now? You do NOT have a blank check to roll-back decades of progress. #WomensMarch We'll hold you accountable • #WomensMarch BECAUSE THIS NATION IS UNDER CONTROL OF A MISOGYNISTIC PIG!
	$N = 13,242$	$N = 3,723$
Opposing stance	• Liberal protests only serve to strengthen the resolve of real Americans that foot the bill. #WomensMarch #MAGA • I feel so blessed to live in the greatest country on earth where I can run my own business. I do not feel a need to protest. #WomensMarch	• This is gross and classless. I'm so sad for how brainwashed our young women are. #WomensMarch • @womensmarch. Congratulations losers, you walked, I hope today you have sore feet, at least you will have accomplished something!!
	$N = 494$	$N = 2,153$
Total	$N = 19,612$	$r = 0.44$

indicating respect for participants) or negative sentiment (expressing fear that rights will be undermined or progress undone, indicating anger about the results of the election). Likewise, opposition to the March can have an overall negative sentiment (criticizing the movement or its supporters) or positive sentiment (indicating satisfaction with the status quo). It is evident from these examples that while support is largely expressed using positive language and vice versa, if we were to simply take the sentiment of tweets about the Women's March as a proxy for the stance of the authors, we would miss the many instances where this is not the case and risk producing a biased measure.

Given the fact that sentiment and stance appear to be only weakly correlated in tweets about the Women's March, it becomes relevant to question whether a more accurate measure of stance might have changed the conclusions Felmlee et al. (2020) arrived at regarding the overall level of approval for the movement contained within the tweets they analyzed. To examine this question, we used a neural network classifier built on top of BERT ("Bidirectional Encoder Representations for Transformers"), a massive pretrained language representation model (Devlin et al. 2018) that represents the current state of the art in language modeling.[4] We trained this classifier on the

4 BERT is a transformer, a recently-introduced neural network architecture for processing sequences of data such as text (Vaswani et al. 2017). While earlier models designed for this task relied on recurrent architectures such as LSTMs to capture term context, recurrence is computationally expensive in neural networks because it requires the sequence to be processed in order, increasing both memory requirements and training time. Transformers rely instead on parallelizable attention mechanisms to establish bidirectional context, making them much more efficient and able to be trained on massive

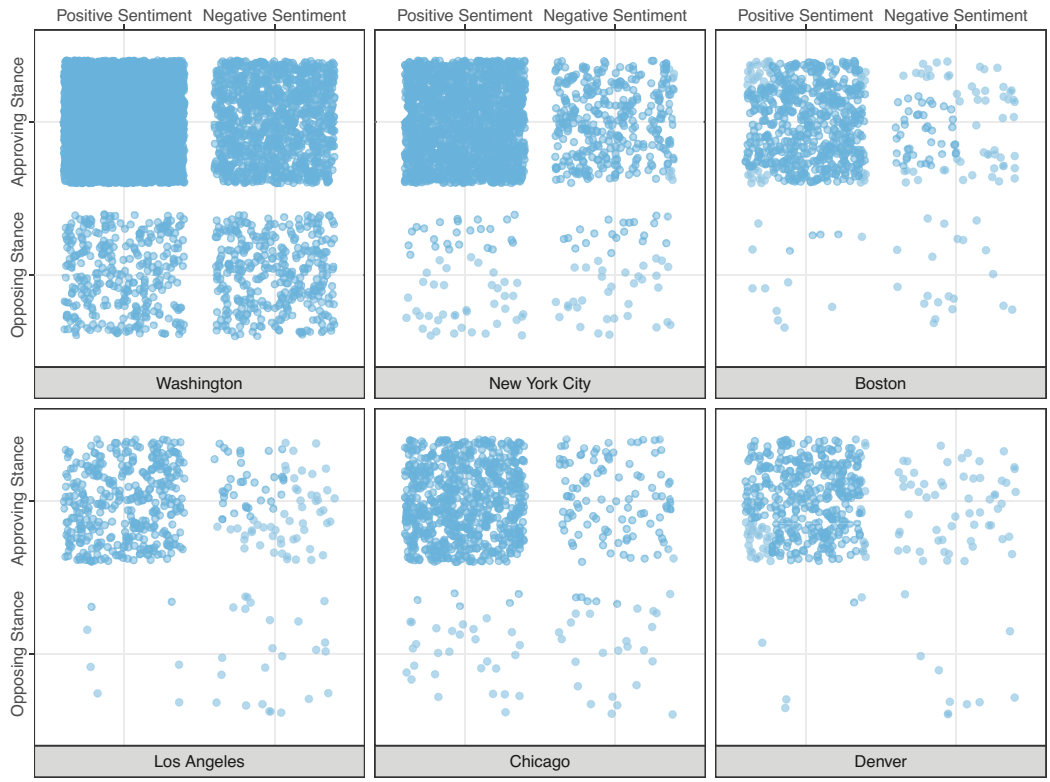

Figure 2. Geographic distribution of machine-coded sentiment and stance in tweets about the Women's March.

ground-truth stance labels for each document in the human-coded Women's March tweets corpus described above, using random undersampling to account for the class imbalance in the training data. The resulting model achieved an average F1-score of 0.853 on a held-out test sample. We then apply this classifier to the replication data from Felmlee et al. (2020), while also dichotomizing their reported VADER sentiment scores as indicating overall positive sentiment (scores above zero) and overall negative sentiment (scores below zero). Tweets that contained no sentiment information at all were dropped. Figure 2 shows the distribution of these modified measures, focusing once again on tweets mentioning the six most-referenced March locations. It quickly becomes apparent that while the above-average positivity of the VADER-generated sentiment scores does indeed reflect the fact that more tweets indicate approval of the Women's March than opposition, these scores actually under-represent the true level of support for the movement expressed throughout the corpus by a fairly substantial amount. In fact, when stance is more accurately measured, approval for the movement is so overwhelming that a tweet employing negative language to discuss the March is actually more likely to be expressing an approving stance than an opposing one.[5]

While Felmlee et al. (2020) are primarily focused on measuring and describing the opinions expressed in tweets about the Women's March, researchers conducting similar studies are often

corpora. Using the transformer architecture, models such as BERT, GPT-2, or XLNet are pretrained to perform masked-term prediction on billions of terms, and in doing so learn contextual word embeddings that appear to capture linguistic structure, and have proven useful for a wide variety of NLP tasks. For our application, we use the Transformers and Simple Transformers libraries in Python (Rajapakse 2020; Wolf et al. 2020) to implement the pretrained BERT-base model with a classifier "head," essentially an additional linear layer on top of the BERT-base encoder.

5 Felmlee et al. (2020) note in their analysis that this particular set of tweets, which all include hashtags referencing specific Women's March locations, may be significantly more supportive than messages about the movement in general, since location-related hashtags are generally used by participants at a particular event, while critics would tend to direct their comments towards the movement at large.

Table 2. Regression analysis: predicting Women's March approval with ideology.

	Women's March approval		
	Machine-coded		Human-coded
	VADER (Sent.)	BERT (Stance)	
	(1)	(2)	(3)
Ideology (lib-cons)	−0.35	−0.76	−1.87
	(0.06)	(0.06)	(0.12)
Constant	0.84	1.24	2.62
	(0.08)	(0.07)	(0.16)
Observations	928	1,501	1,501
Log likelihood	−516.22	−606.61	−256.94
Akaike Inf. Crit.	1,036.44	1,217.23	517.88

further interested in testing hypotheses about how what people think and say on a particular political issue relates to their other beliefs, behaviors, and characteristics. For example, we might want to evaluate the expectation that people who hold a conservative political ideology would be less likely to have a favorable view of the Women's March (a movement protesting against an incoming Republican administration). A common approach to addressing this type of question would be to obtain separate measures of both general political ideology and opinions on the specific issue of interest, then regress the opinion measure on each subject's ideology score. The hypothesis would then be supported if general ideology appears to be a strong predictor of issue opinion. In this type of research design, how well the opinion measure used in the regression analysis actually captures the underlying opinion of interest might matter quite a bit for the inferences we draw from the results, and using a measure of sentiment when what we are actually interested in is a specific stance toward a particular target can create issues of validity, as we illustrate through an extension of the Felmlee et al. (2020) analysis.

We began by taking a new sample from the Women's March corpus containing 1,500 tweets posted by distinct users, and scored each author on an ideological scale ranging form −2.5 (very liberal) to 2.5 (very conservative) according to the Bayesian ideal point estimation approach suggested and validated in Barberá (2015). Barberá's (2015) method leverages the social network structure of Twitter, inferring the latent political preferences of individual Twitter users from the known preferences of elites that user follows. With an ideology measure available, we next employed both the dichotimized VADER classifier and our fine-tuned BERT classifier to extract a measure of opinion on the Women's March from each tweet in the sample. Finally, we regressed both machine-generated opinion measures as well as human-coded ground truth stance labels on each user's ideology score in three simple logistic regression models, shown in Table 2.

We see that in the model with a human-coded dependent variable, representing ground truth stance towards the Women's March, there is a strong association between liberal ideology and approval of the movement, as hypothesized. Both of the models estimated using a machine-coded opinion measure capture the direction of this association, in keeping with our finding that sentiment and stance are somewhat correlated in this corpus. However, the coefficient for ideology in the VADER model is substantially smaller than the equivalent coefficient in the BERT model. Figure 3 shows predicted probabilities generated from all three models, and illustrates that even though the general inferences we might draw from the VADER opinion measures are substantively similar to those generated by the BERT model, the VADER model suggests a weaker relationship

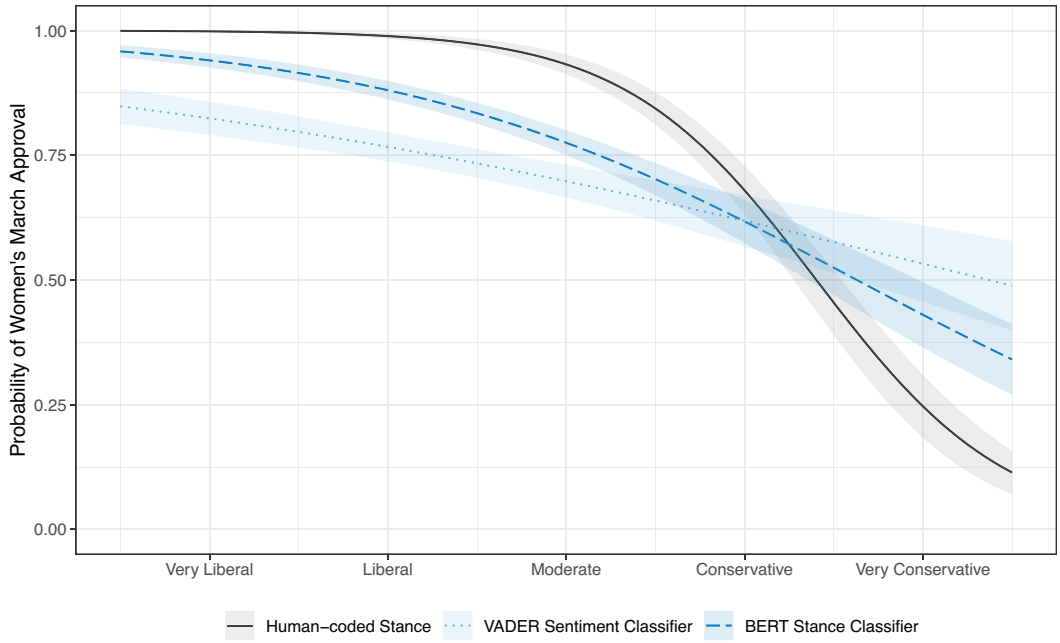

Figure 3. Simulated probabilities: predicting Women's March approval with ideology.

between ideology and opinion and particularly underrates how liberal authors approve of the movement and overrates how conservative authors oppose it. While this is a relatively minor misstep, it might lead us to conclude that the Women's March movement enjoyed a wider base of support across the range of political ideology than was actually the case.

3 Do More Accurate Sentiment Classifiers Make Better Stance Predictions?

The attentive reader might view the replication exercise presented above and rightly object that we have compared the classifications generated by VADER, a relatively simple dictionary-based sentiment analysis method, to those generated by a purpose-tuned BERT model, an approach that represents the current state of the art in NLP. Not only was our BERT classifier fine tuned on the specific task we wanted it to perform, but it also has the benefit of pretraining on massive unstructured corpora, allowing it to leverage sophisticated contextual language representations. There is also a general intuition that lexicon-based sentiment analysis techniques are inevitably less accurate than any purpose-trained supervised classifier, a skepticism which has merit. González-Bailón and Paltoglou (2015) applied an array of sentiment dictionaries to political corpora from a variety of sources ranging from BBC news coverage, to blog posts, to social media, and found that a supervised classifier consistently outperformed them. Additionally, they found that agreement across sentiment measures produced by various dictionary methods is relatively low—particularly when the domain of the corpus is highly specialized. In a sense, stance detection can be thought of as a form of sentiment analysis that is not only specialized to a specific domain, but also to a specific target, so it should be unsurprising that supervised learning is the dominant approach in NLP research on this task. In a recent survey of the stance detection literature, Küçük and Can (2020) found that traditional feature-based machine learning approaches such as support vector machines (SVMs), logistic regression, and naive Bayes classifiers are frequently used, and SVMs in particular commonly serve as the baseline classifier against which new approaches are compared. Deep learning approaches also appear frequently throughout this literature, and often compare favorably to baseline models such as SVMs. LSTMs and other forms of recurrent neural networks

(RNNs) are still most common, but have begun to be replaced by transfer learning approaches with pretrained Transformer models like BERT, which substantially outperform other approaches in stance detection and related tasks (Attardi and Slovikovskaya 2020; Hardalov et al. 2021).

Transformer-based models likely perform as well as they do at stance detection because of their superior ability to deal with language contextually. The meaning of a word is dependent on its context, and this can be of particular relevance in stance detection. The term *love*, for example, suggests one stance in a sentence like "I love all these protesters" and an entirely different one in a sentence like "I'd love it if all these protesters got arrested." Classifiers that use a "bag-of-words" approach (including both lexicons and traditional feature-based classifiers) can only assign one weight to a given term, so in this example, the word *love* would not be a particularly discriminating feature to those classifiers even though it carries significant meaning to a human reader. The transformer architecture, by contrast, uses a self-attention mechanism to relate the terms in a given document to each other, capturing variations in context and providing the classifier with more information that can be used to make a determination between classes.[6]

In the analysis of the Women's March tweets it would have been highly surprising if a state-of-the-art supervised model like our BERT classifier had not outperformed a dictionary-based approach like VADER. However, the fundamental point we are trying to make is not just that text classifiers can vary in accuracy, nor even that a specialized supervised model will generally outperform an off-the-shelf unsupervised approach. Since sentiment is conceptually distinct from stance, models trained to identify a document's sentiment often produce measures that serve as poor proxies for stance, regardless of how accurately they can perform the task for which they were initially developed. In the sections that follow, we demonstrate this fact through two additional examples, where we compare the performance of a wider range of classifiers trained on both ground-truth sentiment and stance information, before once again using the measures produced as the outcome variable in a downstream regression analysis relating authors' general political ideology to the specific opinions they express in the texts. The conceptual distinction between sentiment and stance should cause us to observe sentiment-trained models performing worse on stance identification tasks than models trained on stance. Likewise, in downstream analyses, we should expect to observe different relationships between ideology and opinion measures generated by stance-trained and sentiment-trained models respectively.

3.1 Candidate Methods

The following sections will examine opinions expressed in a corpus of short open-ended survey responses as well as on social media. In these examples, we consider four methods that are commonly used for sentiment analysis: two dictionary-based methods and two supervised classifiers. Together, these represent a range of common approaches and levels of methodological sophistication.

In addition to the VADER dictionary, we also employ the Lexicoder Sentiment Dictionary (LSD), a lexicon for dictionary-based sentiment analysis specifically designed for the analysis of political texts (Young and Soroka 2012). The LSD contains 4,567 positive and negative terms, including 1,021 terms particular to political discussion that do not occur in other lexicons. We apply Lexicoder to each example corpus using its implementation in the R package Quanteda (Benoit et al. 2018). For each document, Lexicoder returns a value equal to the number of negative term matches subtracted from the number of positive term matches. If this value is positive, the document is coded as having an overall positive sentiment, and vice versa. While Lexicoder is specifically tuned for the analysis of political texts, it tends to perform best on longer documents that largely adhere

6 For a more detailed presentation of the self-attention mechanism at the heart of the Transformer architecture, see (Vaswani et al. 2017)

to conventional spelling. The original application of the method, presented in Young and Soroka (2012) was concerned with the sentiment of news articles, for example. Since this is the case, it will be interesting to compare Lexicoder's performance on shorter, informal documents such as social media posts or short answer survey responses to that of VADER, which was not specifically designed for analyzing political discourse, but does tend to perform well with short texts and informal language.

For supervised methods, the BERT model previously introduced will be compared against a support vector machine classifier, which has long been a workhorse model in supervised text classification and is commonly used as a baseline model in the NLP stance detection literature (Küçük and Can 2020). For each example we evaluate a total of four supervised models, one BERT and one SVM classifier trained on ground-truth sentiment labels for each document and another pair trained on ground-truth stance labels.

3.2 Evaluation Procedure

In the first stage of each example, we evaluate and compare the relative performance of each method at sentiment and stance identification, respectively. This is done with fivefold cross-validation, where the corpus is split into five equal samples, and each one serves as a held-out test set in turn. Having five opportunities to measure each accuracy metric provides us with a sense of the variance of each model's performance over slightly different samples. A point estimate for the metric can then be obtained simply by averaging the scores over all five folds. We evaluate each method according to its average F1 score for each task, which balances precision and recall. The two dictionary-based methods require no additional training, and are simply applied to the test sample of each fold for validation.[7] The two supervised classifiers are trained in turn on the ground-truth stance and sentiment labels of the remaining corpus once the held-out sample has been removed. For the SVM models, the text is preprocessed by removing capitalization and punctuation, then tokenizing each document into unigram tokens and representing each term's incidence using TF-IDF weights. For the BERT models, limited preprocessing is required, so case-folding was the only step taken.

After the model evaluation stage, each model is applied to an additional held-out sample of the relevant example corpus in order to construct a measure of each author's opinion on a specific topic in American politics. As in the earlier extension of the Felmlee et al. (2020) Women's March analysis, these measures are then regressed on a measure of authors' general political ideology in a series of logistic regression models. The corpus of open-ended survey responses contains associated objective measures, so for that analysis ideology is self-reported on a standard 5-point Likert scale by each respondent. In the other example, each author is a Twitter user, and self-reported ideology scores are not available. For these analyses, we again situate each author in ideological space using Bayesian ideal point estimation based on the information provided by the other accounts they follow, as outlined in Barberá (2015).

7 When using sentiment dictionaries, documents that have no term matches in the dictionary, or have an equal number of positive and negative matches, are generally coded as sentiment-neutral. However, all of the documents in each example corpus express either a supportive or opposing stance, so we are treating each analysis as a binary classification task, and the possibility of neutral scores creates a problem for evaluation. We deal with this in three ways: one approach is to break ties randomly, effectively forcing the method to make a "guess" instead of returning a neutral score. The F1 scores reported in the main text of the paper are calculated using this approach. A second approach is to simply drop documents coded as neutral when evaluating the method. This inflates the resulting performance metric, since the method is only being evaluated on its confident predictions. Alternatively, neutral predictions can be simply treated as incorrect when calculating performance metrics, which deflates the resulting score. Metrics calculated according to these other approaches can be found in the Supplementary Appendix.

4 Example II: Sentiment and Stance in Open-Ended Survey Responses About Donald Trump

Another common situation where researchers might want to extract opinion measures from text data is in the analysis of open-ended responses on surveys. For example, YouGov and the McCourtney Institute of Democracy at Penn State regularly field the Mood of the Nation poll (MOTN), a nationally-representative survey that focuses on collecting and analyzing free-form text responses (The McCourtney Institute for Democracy 2020). In addition to collecting standard continuous and categorical measures of political and demographic characteristics of survey respondents, the Mood of the Nation also asks a series of open-ended questions, allowing respondents to express their opinions on American politics using their own words. Some questions vary from wave to wave in response to current events, but every wave asks at least four core questions, prompting respondents to discuss what about American politics makes them feel *proud*, *hopeful*, *angry*, and *worried*. As might be expected, since 2016 respondents have frequently referenced Donald Trump in these responses, expressing support or opposition toward him in a variety of ways. The questions themselves are sentiment-laden by construction, but just because a respondent mentions Trump when talking about what makes them proud does not necessarily mean that they are proud of Trump. Table 3 contains several examples that show how in talking about what makes them proud or hopeful, respondents can easily indicate stances of either approval or opposition toward Trump. As with the Women's March tweets example, many documents in this corpus are on the on-diagonal of aligned sentiment and stance, but plenty of misaligned examples exist as well, and the sentiment and stance labels are only correlated at $r = 0.51$.

The corpus used in this example is drawn from nine MOTN waves, conducted between November 16, 2016 and September 26, 2019. We pool survey waves and questions to generate a corpus

Table 3. Sentiment and stance in mood of the nation responses.

	Positive sentiment (Proud/Hopeful)	Negative sentiment (Angry/Worried)
Approving stance	• [I'm proud of] the recent election of Donald Trump and the freedom of the people to choose who represents them • Trump's European trip has made me exceptionally proud. . . . he made America look great	• [I'm angry about] the way the Democrats are fighting our President. . . and wanting to drive this country to Socialism • [I worry about] if a liberal democrat is elected president and all the things Trump has fixed will go back
	$N = 2,087$	$N = 747$
Opposing stance	• [I'm proud of] the democrats standing up to Trump • [I'm hopeful that] the president may actually face impeachment	• [I'm angry that] a despicable and evil clown got elected president. A sad day for America and the world • [I worry that] Donald Trump is ruining this country
	$N = 967$	$N = 3,345$
Total	$N = 7,146$	$r = 0.51$

Table 4. Classifier performance: mood of the nation responses.

Classifier	F1 score (predicting sentiment)	F1 score (predicting stance)
Lexicoder	0.668 (0.003)	0.633 (0.005)
VADER	0.664 (0.005)	0.620 (0.008)
SVM (sentiment-trained)	0.831 (0.004)	0.723 (0.004)
BERT (sentiment-trained)	0.875 (0.002)	0.724 (0.005)
SVM (stance-trained)	0.724 (0.005)	0.817 (0.004)
BERT (stance-trained)	0.741 (0.005)	0.854 (0.003)

Notes: Reported figures are the average F1 score over fivefold cross validation. Standard errors in parentheses.

of 35,842 documents, where each document is a response to one of the four core open-ended questions ("what about American politics makes you feel *proud*, *hopeful*, *angry*, and *worried*"). Then, we subset the corpus to just documents that mention Donald Trump, for a final *N* of 7,146. These documents are then automatically labeled for sentiment and stance. Since these questions are sentiment-laden by construction, we leverage this information to create sentiment labels, inferring positive sentiment for responses to the *proud* or *hopeful* question and negative sentiment for responses to the *angry* or *worried* question. Likewise, since each document is associated with a respondent's answers to other closed-form responses, we create labels for supportive or opposing stance toward Donald Trump for each respondent from objective measures of approval contained in the survey.

4.1 Measuring Presidential Approval from Open-Ended Survey Responses

To further quantify the misalignment of sentiment and stance in the Mood of the Nation short answer corpus, we employ a series of unsupervised and supervised sentiment analysis methods. With each method, we first predict that sentiment of each tweet and evaluate the method's relative performance against the ground truth sentiment values indicated by the question construction, as described above. Then, mimicking the practice of proxying document stance through sentiment, we evaluate the method's relative performance against the ground truth stance values and compare. Finally, for supervised methods, we then retrain the model on ground truth document stance and evaluate the performance improvements achieved through the use of conceptually accurate training data. Each configuration of method, training data, and ground truth test labels is evaluated through fivefold cross validation. Relative performance is reported as F1 score in Table 4.

The first row of Table 4 shows the performance of the LSD for predicting ground truth sentiment and stance in the MOTN corpus. Averaged over the five "folds," Lexicoder achieved an average F1 score of 0.668 when applied to the true sentiment analysis task, but only 0.633 when used

as a proxy for stance. VADER performs about the same as Lexicoder, with an F1 score of 0.664 when predicting sentiment labels, which fell to 0.620 when used to predict stance. Supervised models did better, with SVM scoring 0.831 and BERT scoring 0.875 on average at the sentiment classification task, but those figures fall to 0.723 and 0.724, respectively when the sentiment-trained models are applied to the stance classification task. Models actually trained on stance labels do the best at the stance identification task, with average F1 scores of 0.817 and 0.854, respectively. Additionally, as might be expected, these stance-trained models are less accurate at the sentiment identification task. Overall, we see a drop in F1 score of about 9 points on average when a model trained on one set of ground-truth labels is used to predict the other.

This brief exercise demonstrates a number of things. First, and perhaps unsurprisingly, sentiment analysis techniques work reasonably well at measuring sentiment. Off-the-shelf dictionary methods Lexicoder and VADER both perform better than random, although they were both substantially outperformed by the supervised methods we tried. If labeled training data is available, supervised models generally make better text classifiers, and transfer learning approaches with pretrained transformers like BERT clearly represent a step forward in supervised text classification. However, increasingly sophisticated models cannot solve the fact that sentiment is a fundamentally imperfect proxy for stance. Even in this corpus where the correlation between sentiment and stance is relatively high, all four approaches considered had substantial drops in performance when tested on ground truth stance instead of sentiment.

4.2 Relating Political Ideology to Presidential Approval

The second step of this example is to employ the opinion measures produced by the classifiers examined above in a related downstream regression analysis, as we did in our extension of the Felmlee et al. (2020) analysis. We again take a held-out sample from the MOTN corpus, and use the classifiers from the previous section to produce seven different dichotomous measures (one from each of the six classifiers as well as the self-reported approval measure) indicating whether each respondent in the sample approves or disapproves of Donald Trump. Then, in a series of logistic regression models, we regress these opinion measures on an objective measure of general political ideology recorded in the MOTN survey to test the expectation that more conservative respondents will be more likely to support Trump, a Republican president, while liberal respondents will be more likely to oppose him. These results appear in Table 5.

The expectation is that more conservative respondents will be more likely to support Trump, so we would expect to see positive coefficients on political ideology in models where the support measure used is produced by a classifier that is capturing true stance information from the text. This is what we see for all seven models, consistent with our finding that sentiment and stance are weakly correlated in this corpus as well. But as in the Women's March example, coefficients for ideology are substantially smaller in regression models where the measure of Trump approval used as the outcome variable is actually capturing general sentiment. The models with machine-coded outcome variables that come the closest to replicating the true relationship between general political ideology and self-reported approval of Trump shown in model 7 are models 5 and 6, where the outcome variable is produced by a stance-trained classifier.

Incorrectly proxying stance with sentiment has somewhat more dramatic effects on the inferences we might draw from the regression models in this example. Figure 4 shows the predicted probability curves produced by the models associated with the most accurate sentiment and stance-trained models (models 4 and 6, respectively) as well as the self-reported model as ideology moves from very liberal to very conservative. As in the Women's March example, both models using machine-generated measures capture the expected direction of the relationship, but in the sentiment-trained model, Trump approval among liberals is being significantly overestimated

https://doi.org/10.1017/pan.2022.10 Published online by Cambridge University Press

Table 5. Regression analysis: predicting trump approval with ideology.

	Trump approval						
	Machine-coded						Self-reported
	Lexicoder	VADER	SVM (Sent.)	BERT (Sent.)	SVM (Stance)	BERT (Stance)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Ideology (lib-cons)	0.25	0.24	0.39	0.41	0.81	0.99	2.20
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.05)
Constant	−0.84	−0.77	−1.65	−1.65	−3.12	−3.61	−7.22
	(0.11)	(0.10)	(0.10)	(0.10)	(0.12)	(0.13)	(0.20)
Observations	5,892	5,991	7,146	7,146	7,146	7,146	7,146
Log Likelihood	−3,964.94	−4,067.80	−4,529.57	−4,546.55	−3,880.06	−3,684.33	−2,396.70
Akaike Inf. Crit.	7,949.89	8,155.61	9,079.13	9,113.09	7,780.11	7,388.65	4,813.41

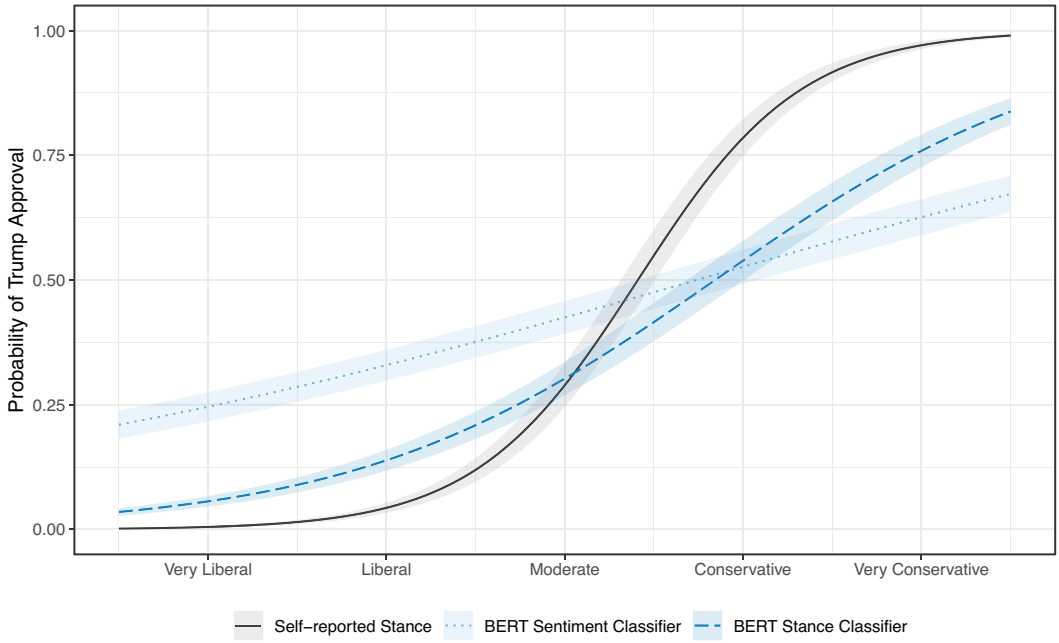

Figure 4. Simulated probabilities: predicting trump approval with ideology.

compared to the stance-trained model, and although the stance-trained model is closer, neither quite captures the strength of Trump approval among very conservative respondents.

5 Example III: Sentiment and Stance in Tweets About the Kavanaugh Confirmation

In both the Women's March example and the Mood of the Nation example, sentiment and stance were at least correlated in the training corpus. In this final example, we explore what happens when sentiment is used as a proxy for stance in a corpus where the two concepts are not only misaligned, but are essentially orthogonal. In the Fall of 2018, Brett Kavanaugh was nominated to a seat on the U.S. Supreme Court. While the Senate was considering the nomination, a university professor, Christine Blasey Ford, stepped forward claiming that Kavanaugh had sexually assaulted her at a party when they were both teenagers. Ford agreed to testify to Congress in the confirmation hearings, which quickly became highly contentious and partisan. The hearings attracted a great deal of public attention on the national stage, and public discourse from this time reflects the high levels of (largely negative) emotions that defined the proceedings. While the hearings were underway, Baumgartner (2018) collected a corpus of over 50 million tweets about Kavanaugh, the confirmation process, and the assault allegations, which we sample for this example and hand-code for sentiment and stance values as in the Women's March example. Table 6 shows a selection of tweets from this corpus, illustrating that, as with the examples discussed earlier, it is possible for both approving and opposing stances towards the confirmation of Brett Kavanaugh to be expressed using either positive or negative sentiment.

Approval of Kavanaugh's confirmation can be expressed by speaking positively about the man or his qualifications, or by speaking negatively about the process, allegations, or delays. Likewise, opposition to the confirmation can be expressed by speaking positively about Ford, or negatively about Kavanaugh and his supporters. For the most part, however, discourse on both sides of the issue was largely characterized by negative sentiment. People who supported the confirmation were upset that the process was being delayed by allegations they viewed as irrelevant and/or specious. People who opposed it viewed the allegations as both credible and

Table 6. Sentiment and stance in tweets about the Kavanaugh hearings.

	Positive sentiment	Negative sentiment
Approving stance	• #ConfirmKavanaugh this is a great man he will make a great Supreme Court judge • POWERFUL TESTIMONY of women that stand behind Brett Kavanaugh. Calling him Honorable with High Integrity, Family man, great friend and boss!	• The Republicans can certainly kiss the midterms goodbye if they blow this Kavanaugh confirmation • The sinister left will do absolutely anything to maintain power and attack conservatives ....Lie, cheat, delay. Whatever it takes. We absolutely cannot allow them to win. Confirm Kavanaugh!
	$N = 521$	$N = 1,467$
Opposing stance	• I hope she feels the love and support and the heartache that women feel in standing in solidarity with her • DR. FORD IS AN AMERICAN HERO	• Kavanaugh and the GOP have no idea of the power and anger they are unleashing • @SenateGOP withdraw Kavanaugh, you are just dragging yourselves, this country and Dr. Ford through the mud
	$N = 391$	$N = 1,281$
Total	$N = 3,660$	$r = 0.03$

serious, and were deeply troubled by the prospect of someone with a history of sexual assault sitting on the Supreme Court. Table 6 also shows the distribution of positive and negative tweets indicating approving and opposing stance in our sample, and although positive sentiment does appear in the corpus, it is relatively rare, and sentiment and stance are essentially uncorrelated ($r = 0.03$).

5.1 Measuring Opinion from Tweets About the Kavanaugh Confirmation Hearings

As with the Mood of the Nation example, we employ a series of unsupervised and supervised sentiment analysis methods in order to help quantify the misalignment of sentiment and stance in the Kavanaugh tweets corpus. Relative performance over fivefold cross-validation is reported as average F1 score in Table 7.

Table 7 illustrates the difference between sentiment and stance in this corpus quite dramatically. At the sentiment classification task, each of the sentiment-trained methods performs about as well or better than it did in the previous example. Lexicoder achieved an average F1 score of 0.788, VADER a 0.754, sentiment-trained SVM a 0.943 and sentiment-trained BERT a 0.954. However, when applied to the task of identifying author stance for these documents, none of these methods could do much better than random, displaying a drop in F1 score of over 30 points on average between the two tasks. Likewise, the supervised models trained on stance labels only

Table 7. Classifier performance: Kavanaugh tweets.

Classifier	F1 Score (predicting sentiment)	F1 Score (predicting stance)
Lexicoder	0.788	0.572
	(0.005)	(0.014)
VADER	0.754	0.514
	(0.005)	(0.011)
SVM (sentiment-trained)	0.943	0.514
	(0.003)	(0.012)
BERT (sentiment-trained)	0.954	0.582
	(0.002)	(0.005)
SVM (stance-trained)	0.584	0.935
	(0.007)	(0.006)
BERT (stance-trained)	0.576	0.938
	(0.008)	(0.002)

Notes: Reported figures are the average F1 score over fivefold cross validation. Standard errors in parentheses.

barely outperform a random baseline when applied to the task of identifying sentiment, while when appropriately applied to the stance detection task they again achieved high F1 scores of 0.935 (SVM) and 0.938 (BERT).

5.2 Relating Political Ideology to Kavanaugh Opinions

When we use these classifiers to produce opinion measures on a held-out sample for a downstream analysis, the pattern is even more defined. We again use the Barberá's (2015) Bayesian ideal point estimation method to assign ideology scores to the Twitter users who wrote the tweets in the sample. The general expectation in this analysis is that conservatives are going to be more likely to approve of the confirmation of Kavanaugh, a conservative judge nominated by a Republican president. Therefore, if our classifiers are accurately capturing stance, then when we regress Kavanaugh approval on an ideology scale that ranges from liberal to conservative, we should see large positive coefficients for political ideology. Table 8 shows that this is not what we observe. Out of the four models using opinion measures generated by sentiment-trained classifiers, only one estimates a coefficient that is statistically distinguishable from zero at conventional thresholds for significance, and it remains quite small in magnitude. By contrast, the two models that use opinion measures generated by stance-trained classifiers have large, positive, and statistically significant coefficients for ideology, coming much closer to reproducing the strong relationship between author conservatism and approval of Kavanaugh's confirmation that is visible in model 7, where the outcome variable is human-coded stance toward Kavanaugh.

This extreme attenuation of coefficients would have a large effect on the inference we would draw from this analysis, were we to try to capture support for Kavanaugh by proxying stance with sentiment. Figure 5 shows the predicted probability curves produced by the most accurate sentiment and stance-trained classifier from the training and evaluation phase (once again, this is the BERT classifier in both cases, models 4 and 6 in the regression analysis) along with the predicted probability curve generated by the ground truth model (model 7). We see that as

https://doi.org/10.1017/pan.2022.10 Published online by Cambridge University Press

Table 8. Regression analysis: predicting Kavanaugh approval with ideology.

	Kavanaugh approval						
	Machine-coded						Human-coded
	Lexicoder	VADER	SVM (Sent.)	BERT (Sent.)	SVM (Stance)	BERT (Stance)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Ideology (lib-cons)	0.04	0.01	0.09	−0.005	1.32	1.37	2.16
	(0.03)	(0.02)	(0.03)	(0.03)	(0.04)	(0.04)	(0.08)
Constant	−1.03	−0.41	−1.41	−1.13	−0.63	−0.63	−1.48
	(0.05)	(0.04)	(0.05)	(0.04)	(0.07)	(0.07)	(0.12)
Observations	2,582	2,708	2,785	2,785	2,785	2,785	2,785
Log likelihood	−1,494.56	−1,821.96	−1,390.25	−1,545.72	−962.23	−924.56	−529.36
Akaike Inf. Crit.	2,993.13	3,647.92	2,784.50	3,095.43	1,928.47	1,853.13	1,062.72

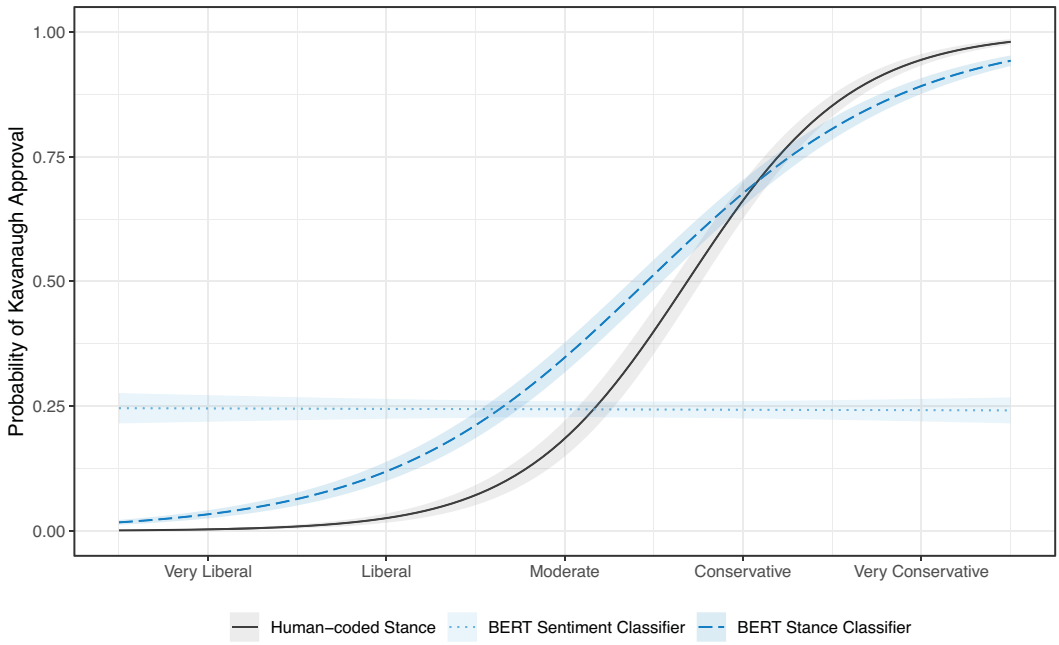

Figure 5. Simulated probabilities: predicting Kavanaugh approval with ideology.

author ideology moves from very liberal to very conservative, the predicted probability of that author approving of Kavanaugh's confirmation (as measured by the stance-trained BERT classifier) increases dramatically, closely mirroring the ground truth curve. However, when the opinion measure used is produced by the sentiment-trained BERT classifier, the relationship between ideology and Kavanaugh approval disappears altogether. If we were to adopt this method in an analysis relating political ideology to support for the confirmation, we would draw the incorrect conclusion that Kavanaugh was uniformly disliked and that regardless of political persuasion, very few Americans wanted to see him seated on the Supreme Court.

6 Discussion and Conclusions

The findings from the examples presented above can be condensed into four major points. First, hopefully we have argued convincingly that researchers in political science should adopt the growing convention in NLP that sentiment and stance are distinct concepts that, although related, cannot necessarily be treated as interchangeable. This distinction is of particular relevance for the analysis of political texts, because political opinions are typically complex and multidimensional enough that it is trivial to express them either negatively or positively. Second, although there can be substantial variation in the accuracy of different approaches to text classifications, the source of the difficulty in getting an accurate measure of stance from a corpus using sentiment analysis techniques is not the relative complexity of the technique chosen. From simple dictionary-based term-matching methods to state-of-the-art neural network transfer learning approaches, all of the sentiment-trained classifiers evaluated in each of the examples considered performed worse at stance identification than at sentiment identification, and all of the stance-trained models performed better at stance identification than their sentiment-trained counterparts did. Any method is going to produce noisy, inaccurate measures if it is trained on labels that reflect something conceptually distinct from the true quantity of interest. Third, using these noisy, inaccurate measures in downstream analyses can produce biased or attenuated results that fail to capture the true relationship between a political opinion and some other variable of interest, and can lead researchers to draw incorrect conclusions. This is a point that can sometimes be lost in the NLP literature, which tends to evaluate methods by comparing

their accuracy on benchmark datasets. While understanding differences in model performance is certainly informative, in many studies using political texts, the goal is not simply to accurately measure a concept of interest from the documents in a corpus, it is to then use that measure to better understand a political process or relationship, and relatively small differences in model performance can have dramatic impacts on the inferences we draw. Finally, the amount of measurement bias introduced by the practice of proxying stance through sentiment is a function of the correlation between ground truth sentiment and stance in the corpus. In some applications, these quantities will be highly correlated, and the worst result of conflating the two will be a small drop in classifier accuracy and mildly attenuated regression coefficients that still capture the overall direction of the true relationship. In other applications, sentiment and stance can be completely orthogonal, and treating them as equivalent can have serious effects on the inferences we make.

These empirical observations suggest some practical advice for political science practitioners who employ text-as-data methods in their research. To begin with, before making use of sentiment dictionaries, it is a good idea to give some thought to whether or not sentiment is truly the quantity of interest. If the focus of the research is on the overall tone exhibited in the text (as it might be in studies of news coverage, campaigns, or toxicity in online communications, for example) then a dictionary approach might be sufficiently effective. Nevertheless, a supervised classifier will almost always outperform dictionary methods, given enough training data, so the best practice is typically going to be to train a domain-specific model to directly identify the outcome of interest. This is a general point that extends beyond the misapplication of sentiment: when working with text data, if there is a particular quantity of interest to be extracted from a corpus, hand-labeling a sample of documents for that quantity, then using it to train a new supervised classifier is usually going to be more effective than relying on an existing model or dictionary designed to identify a different concept, even if it's a related one. Note, for example, that in the analyses presented above, stance-trained classifiers performed about as poorly at sentiment identification tasks as the reverse. Recent advancements in transfer learning have made training a new supervised classifier a less time-consuming prospect than it once was, as models like BERT benefit from massive corpora at the pre-training stage, allowing them to be fine-tuned to perform specific tasks fairly accurately with relatively little additional training data. In cases where an existing model or dictionary is used anyway, it is worth spending the time to hand label a small validation sample at the very least in order to ascertain the correlation between the machine-generated labels and the true concept of interest and determine if any severe misalignments have occurred.

Sentiment analysis has been an attractive approach to extracting opinions from text because it is conceptually tractable and, thanks to the efforts of researchers developing sentiment lexicons, very easily implemented using off-the-shelf tools. Unfortunately, for many applications in the analysis of political texts, sentiment is just not the concept that researchers are interested in capturing. We are often less concerned with the overall polarity of the document and more interested in the particular attitudinal stance expressed in that document. While sentiment can often be correlated with stance in political texts, in this paper, we have demonstrated that one does not necessarily make a good proxy for the other. Models trained on sentiment labels or provided with sentiment lexicons experience substantial reductions in accuracy when asked to predict document stance. Measures of stance produced by these models are noisy and produce biased coefficient estimates when used in downstream regression analyses. Simply put, sentiment is not the same thing as stance, and treating them as conceptually interchangeable can introduce significant measurement error. The fact that stance is a conceptually target-dependent concept makes stance identification a more difficult, context-specific classification task that requires target-aware training data.

Acknowledgments

We are indebted to Diane Felmlee, Justine Blanford, Stephen Matthews, and Alan MacEachren for graciously providing us with access to their data and replication materials. We would also like to thank Bruce Desmarais, Sangyeon Kim, Mikaela Karstens, Howard Liu, Claire Kelling, and Medha Uppala at the Penn State Center for Social Data Analytics for feedback on an earlier draft of this article, as well as the participants of the Text Analysis Across Domains (TextXD) 2019 conference for helpful comments. We also thank the editorial team at *Political Analysis* and three anonymous reviewers for their detailed engagement and suggestions.

Funding

Funding for this project was provided by the Center for Social Data Analytics at Pennsylvania State University.

Data Availability Statement

Data and replication materials for this article are available at Bestvater and Monroe (2022) at https://doi.org/10.7910/DVN/MUYYG4.

Supplementary Material

For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2022.10.

References

Abercrombie, G., F. Nanni, R. T. Batista-Navarro, and S. P. Ponzetto. 2019. "Policy Preference Detection in Parliamentary Debate Motions." In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 249–259. Stroudsburg: Association for Computational Linguistics.

Attardi, G., and V. Slovikovskaya. 2020. "Transfer Learning from Transformers to Fake News Challenge Stance Detection (FNC-1) Task." In *Proceedings of the 12th Language Resources and Evaluation Conference*, 1211–1218. Paris: European Language Resources Association.

Barberá, P. 2015. "Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data." *Political Analysis* 23 (1): 76–91.

Baumgartner, J. 2018. *Kavanaugh Twitter Dataset.* https://pushshift.io/kavanaugh-twitter-dataset/.

Benoit, K., et al. 2018. "Quanteda: An R Package for the Quantitative Analysis of Textual Data." *Journal of Open Source Soffware* 3 (30): 774.

Bestvater, S. E., and B. L. Monroe. 2022. Replication Data for: Sentiment is Not Stance: Target-Aware Opinion Classification for Political Text Analysis. V. V1. https://doi.org/10.7910/DVN/MUYYG4.

Biber, D., and E. Finegan. 1988. "Adverbial Stance Types in English." *Discourse Processes* 11 (1): 1–34.

Choy, M., M. Cheong, M. N. Laik, and K. P. Shung. 2012. "US Presidential Election 2012 Prediction Using Census Corrected Twitter Model." Preprint, arXiv:1211.0938.

Choy, M., M. L. F. Cheong, M. N. Laik, and K. P. Shung. 2011. "A Sentiment Analysis of Singapore Presidential Election 2011 Using Twitter Data with Census Correction." Preprint, arXiv:1108.5520.

Dahal, B., S. A. Kumar, and Z. Li. 2019. "Topic Modeling and Sentiment Analysis of Global Climate Change Tweets." *Social Network Analysis and Mining* 9 (1): 24.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. "Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding." Preprint, arXiv:1810.04805.

Du Bois, J. W. 2007. "The Stance Triangle." In *Stancetaking in Discourse: Subjectivity, Evaluation, Interaction*, edited by R. Englebretson, 139–182. Amsterdam: John Benjamins Publishing Company.

Felmlee, D. H., J. I. Blanford, S. A. Matthews, and A. M. MacEachren. 2020. "The Geography of Sentiment Towards the Women's March of 2017." *PLoS One* 15 (6): e0233994.

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

Gurr, T. R. 1970. *Why Men Rebel*. Princeton: Princeton University Press.

Hardalov, M., A. Arora, P. Nakov, and I. Augenstein. 2021. "A Survey on Stance Detection for Mis-and Disinformation Identification." Preprint, arXiv:2103.00242.

Hopp, T., and C. J. Vargo. 2017. "Does Negative Campaign Advertising Stimulate Uncivil Communication on Social Media? Measuring Audience Response Using Big Data." *Computers in Human Behavior* 68: 368–377.

Hutto, C. J., and E. Gilbert. 2014. "*Vader: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text*." In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, 216–225. Palo Alto: Association for the Advancement of Artificial Intelligence.

Jose, R., and V. S. Chooralil. 2015. "Prediction of Election Result by Enhanced Sentiment Analysis on Twitter Data Using Word Sense Disambiguation." In *2015 International Conference on Control Communication & Computing India (ICCC)*, 638–641. Bangalore: IEEE.

Klašnja, M., P. Barberá, N. Beauchamp, J. Nagler, and J. A. Tucker. 2015. "Measuring Public Opinion with Social Media Data." In *The Oxford Handbook of Polling and Polling Methods*, edited by L. R. Atkeson and R. M. Alvarez, 555–582. New York: Oxford University Press.

Küçük, D., and F. Can. 2020. "Stance Detection: A Survey." *ACM Computing Surveys (CSUR)* 53 (1): 1–37.

Maas, A. L., R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. 2011. "Learning Word Vectors for Sentiment Analysis." In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 142–150. Stroudsburg: Association for Computational Linguistics. http://www.aclweb.org/anthology/P11-1015.

Mohammad, S., S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry. 2016. "Semeval-2016 Task 6: Detecting Stance in Tweets." In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 31–41. Stroudsburg: Association for Computational Linguistics.

Murthy, D. 2015. "Twitter and Elections: Are Tweets, Predictive, Reactive, or a Form of Buzz?" *Information, Communication & Society* 18 (7): 816–831.

O'Connor, B., R. Balasubramanyan, B. R. Routledge, and N. A. Smith. 2010. "*From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series*." In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 122–129. Palo Alto: Association for the Advancement of Artificial Intelligence.

Pang, B., L. Lee, and S. Vaithyanathan. 2002. "Thumbs Up? Sentiment Classification Using Machine Learning Techniques." Preprint, arXiv:cs/0205070.

Rajapakse, T. 2020. Simple Transformers. https://simpletransformers.ai/.

Rezapour, R., L. Wang, O. Abdar, and J. Diesner. 2017. "Identifying the Overlap Between Election Result and Candidates' Ranking Based on Hashtag-Enhanced, Lexicon-Based Sentiment Analysis." In *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*, 93–96. New York: IEEE.

Soroka, S., L. Young, and M. Balmas. 2015. "Bad News or Mad News? Sentiment Scoring of Negativity, Fear, and Anger in News Content." *The Annals of the American Academy of Political and Social Science* 659 (1): 108–121.

The McCourtney Institute for Democracy. 2020. The Mood of the Nation Poll. https://democracy.psu.edu/research/mood-of-the-nation-poll/.

Tumasjan, A., T. O. Sprenger, P. G. Sandner, and I. M. Welpe. 2010. "*Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment*." In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 178–185. Palo Alto: Association for the Advancement of Artificial Intelligence.

Vaswani, A., et al. 2017. "Attention is All You Need." In *Advances in Neural Information Processing Systems*, 5998–6008. San Diego: Neural Information Processing Systems Foundation.

Wang, H., D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan. 2012. "A System for Real-Time Twitter Sentiment Analysis of 2012 US Presidential Election Cycle." In *Proceedings of the ACL 2012 System Demonstrations*, 115–120. Stroudsburg: Association for Computational Linguistics.

Wolf, T., et al. 2020. "Transformers: State-of-the-Art Natural Language Processing." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online*, 38–45. Stroudsburg: Association for Computational Linguistics.

Young, L., and S. Soroka. 2012. "Affective News: The Automated Coding of Sentiment in Political Texts." *Political Communication* 29 (2): 205–231.