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

Samuel E. Bestvater* Burt L. Monroe*

Contents

1 Annotation Procedure 2

2 Dictionary Methods and Unlabeled Documents 4

3 Stance and Sentiment Intensity 7

*Dept. of Political Science, Penn State, University Park, PA.

1 Annotation Procedure

Two of the example analyses presented in the main text of the article rely on two corpora of tweets on the topics of the Women's March and Kavanaugh confirmation hearings, respectively, that have been hand-labeled according to their sentiment and stance values. These hand-labeled corpora are available in the online replication packet associated with the article. Details on the annotation process are provided below, along with information about annotator agreement.

The annotation of these corpora was performed by two coders using the annotation platform Labelbox. Coders were shown a target (either the Women's March or the confirmation of Brett Kavanaugh) along with the text of a tweet on that topic. Then they were asked to answer two questions about the text: first, whether the general sentiment of the language used in the text is *positive* or *negative* and second, whether the specific stance the author expresses toward the provided target is *approving* or *opposing*. As the vast majority of tweets in both corpora expressed clear opinions and did so in clearly sentiment-laden language, Coders were not provided with a *neutral* option for either sentiment or stance, and were instead asked to select labels from the *positive/negative* and *approving/opposing* binaries. In keeping with the advice found in Barberá et al. (2021), we maximized the number of unique texts each coder labeled, though as a check on reliability we also selected samples of ∼200 tweets from each corpus to be annotated by both coders. Inter-coder agreement figures are presented in Tables A1-A4 below.

Table A1: Inter-coder Consensus: Sentiment in Women's March Tweets Corpus

				Coder 2	
				Negative	Positive
N	211	Coder 1	Negative	93	13
% Agreement	88		Positive	13	92
Cohen's κ	0.75				

Table A1 shows annotator consensus for coding sentiment in tweets about the Women's March. 211 tweets were double-coded for consensus, and coders 1 and 2 had 88% agreement for that sample, with a Cohen's κ value of 0.75.

Table A2: Inter-coder Consensus: Stance in Women's March Tweets Corpus

				Coder 2	
				Negative	Positive
N	211	Coder 1	Negative	92	10
% Agreement	91		Positive	8	101
Cohen's κ	0.83				

Table A2 shows annotator consensus for coding stance in tweets about the Women's March. The same 211 tweets were double-coded for consensus, and coders 1 and 2 had 91% agreement for that sample, with a Cohen's κ value of 0.83.

Table A3: Inter-coder Consensus: Sentiment in Kavanaugh Tweets Corpus

				Coder 2	
				Negative	Positive
N	200	Coder 1	Negative	151	0
% Agreement	81		Positive	38	11
Cohen's κ	0.30				

Table A3 shows annotator consensus for coding sentiment in tweets about the Kavanaugh confirmation hearings. 200 tweets were double-coded for consensus, and coders 1 and 2 had 81% agreement for that sample, with a Cohen's κ value of 0.30. This low κ value is worthy of note, particularly since % agreement in this sample remains relatively high. This occurs because of the high imbalance in the sentiment of the tweets in this corpus, which are overwhelmingly negative regardless of their stance. Out of the 200 tweets that were double-coded, 151 were labeled as negative by both coders. Of the remaining 49, 11 were labeled as positive by both coders, and 38 were labeled as positive by Coder 1 only. A closer look at the disagreement here is interesting, and illustrates an additional limitation of sentiment analysis that is outside the scope of this article, but has received attention in NLP in the past: humans commonly use sentiment-laden terms sarcastically, and sarcasm can be difficult to identify, for both humans and machine learning systems. For instance, one of the tweets that received different labels from Coder 1 and Coder 2 reads "Congratulations, Brett Kavanaugh! You've just won yourself another investigation." Coder 1 took positive terms like "congratulations" and "won" at face value and labeled this tweet as *positive*, while Coder 2 took the author's clearly sarcastic tone into account and labeled it as *negative*, even though both coders correctly identified that the author's stance towards Kavanaugh's confirmation is *opposing*. A similar issue arises with idiomatic phrases, such as "with all due respect," which appears in another text the coders disagreed on: "With all due respect Senator Grassley, your committee has given Dr. Ford plenty of time to testify." Coder 1 saw the presence of the term "respect" as indicative of positive sentiment, while Coder 2 saw the underlying meaning (*hurry up!*) as indicative of negative sentiment. Again, both coders recognized that this author was expressing an *approving* stance towards Kavanaugh's confirmation.

Table A4: Inter-coder Consensus: Stance in Kavanaugh Tweets Corpus

N	200
% Agreement	98
Cohen's κ	0.95

		Coder 2	
		Negative	Positive
Coder 1	Negative	101	2
	Positive	3	94

Finally, Table A4 shows annotator consensus for coding stance in tweets about the Kavanaugh confirmation hearings. The same 200 tweets were double-coded for consensus, and coders 1 and 2 had 98% agreement for that sample, with a Cohen's κ value of 0.95.

2 Dictionary Methods and Unlabeled Documents

In order to maintain comparability between supervised classifiers and lexicon-based sentiment scoring methods in this analysis, we treated each sentiment and/or stance identification task as a binary classification task and dichotomized the raw sentiment scores produced by VADER and Lexicoder into categorical values of positive (for scores above 0) and negative (for scores below 0). While this process is relatively straightforward, it does create a particular problem: lexicon-based methods generally produce a document-level score by summing up and then normalizing all the term-level sentiment values of the terms in the document. This means that documents that have no term matches in the dictionary generally receive a value of exactly 0 and are considered sentiment-neutral. While it is of course possible for such a situation to arise because a given document truly expresses no sentiment whatsoever, it's more likely that the document does contain some sentiment-laden terms which are simply not recognized by the dictionary supplied. This issue is especially common when applying these techniques to relatively short documents, such as those found in the example corpora in the present analysis, all of which were overwhelmingly found to contain clearly identifiable sentiment polarities as well as expressions of specific attitudinal stances. Because very few true "neutral" examples appeared during the hand labeling process, and the BERT and SVM classifiers all produced binary positive-negative classifications, in order to compare methods we needed a way to deal with cases where the lexicon-based methods were unable to make a prediction for a given document. We arrived at three different approaches: the first is to break ties randomly, effectively forcing the method to make a "guess" instead of returning a score of 0. All of the F1 scores reported for VADER or Lexicoder methods in the main text of the article us this approach. In addition, a

second approach was to drop the documents these methods scored as 0 when calculating F1 scores, and a third is to treat any scores of 0 as simply incorrect when calculating F1 scores. Table A5 shows the category counts for each method and set of ground truth labels. Table A6 shows how employing these alternate approaches affects the results reported in the main text.

Table A5: Label Distribution, True & Predicted

	Mood of the Nation Corpus			**Kavanaugh Tweets Corpus**		
	Negative	*No Score*	*Positive*	*Negative*	*No Score*	*Positive*
Ground Truth Sentiment	4092		3054	2748		912
Ground Truth Stance	4312		2834	1672		1988
Lexicoder	2028	3827	1291	2240	784	636
VADER	1963	3561	1622	1954	264	1442
SVM (sentiment-trained)	4421		2725	2872		788
SVM (stance-trained)	4704		2442	1681		1979
BERT (sentiment-trained)	4353		2793	2804		856
BERT (stance-trained)	4611		2535	1659		2001

Table A6: Classifier Performance

Classifier	Mood of the Nation Corpus			Kavanaugh Tweets Corpus		
	Total Sample	F1 Score (Sent. Preds.)	F1 Score (Stance Preds.)	Total Sample	F1 Score (Sent. Preds.)	F1 Score (Stance Preds.)
Lexicoder (randomize)	7,146	0.668 (0.003)	0.633 (0.005)	3,660	0.788 (0.005)	0.572 (0.014)
Lexicoder (drop)	3,319	0.806 (0.005)	0.709 (0.006)	2,876	0.831 (0.003)	0.589 (0.019)
Lexicoder (strict)	7,146	0.409 (0.005)	0.359 (0.005)	3,660	0.706 (0.005)	0.482 (0.016)
VADER (randomize)	7,146	0.664 (0.005)	0.620 (0.008)	3,660	0.754 (0.005)	0.514 (0.011)
VADER (drop)	3,585	0.778 (0.004)	0.670 (0.011)	3,396	0.766 (0.005)	0.518 (0.012)
VADER (strict)	7,146	0.414 (0.004)	0.353 (0.007)	3,660	0.724 (0.005)	0.482 (0.012)
SVM (sentiment-trained)	7,146	0.831 (0.004)	0.723 (0.004)	3,660	0.943 (0.003)	0.514 (0.012)
BERT (sentiment-trained)	7,146	0.875 (0.002)	0.724 (0.005)	3,660	0.954 (0.002)	0.582 (0.005)
SVM (stance-trained)	7,146	0.724 (0.005)	0.817 (0.004)	3,660	0.584 (0.007)	0.935 (0.006)
BERT (stance-trained)	7,146	0.742 (0.005)	0.854 (0.003)	3,660	0.576 (0.009)	0.938 (0.002)

Randomize: documents with no dictionary matches are assigned a label at random
Drop: documents with no dictionary matches are dropped from consideration
Strict: documents with no dictionary matches are considered incorrect classifications

Reported figures are the average F1 score over 5-fold cross validation
Standard Errors in parentheses

For both example corpora, the overall result is as expected. F1 scores calculated after dropping documents scored as 0 are inflated relative to the metrics produced when 0 scores are dealt with by randomly assigning a label, which makes sense since the method is only being evaluated on its confident predictions. Likewise, when 0 scores are simply considered incorrect predictions, the resulting F1 scores are deflated by comparison. However, the pattern that remains consistent across all three approaches is that when the opinion measures produced by these lexicon-based approaches are compared against ground-truth human-coded stance labels, they perform worse than they did when evaluated against ground-truth sentiment labels.

3 Stance and Sentiment Intensity

One notable difference between lexicon-based and classification approaches to sentiment analysis is that sentiment dictionaries are easily adapted to sentiment intensity scaling, where rather than classifying documents into one of two (*positive-negative*) or three (*positive-neutral-negative*) categories, documents are assigned a numerical score indicating not just the polarity of sentiment, but also the degree to which that sentiment is expressed. Many dictionary implementations in software will actually return such scores as a default, and provide guidelines for how to transform them into categories if needed. VADER, for example, produces scores ranging from -4 to $+4$ which we have translated into classifications to compare directly against other classifiers. Nevertheless, the fact that the raw scores can be interpreted as sentiment intensity prompts an interesting question: does sentiment intensity affect the correlation between general sentiment and stance toward a given target within a corpus? This is a reasonable hypothesis. Moderate sentiment suggests a less emotionally-charged opinion, so perhaps stances expressed in a more level-headed way are more likely to be aligned with the sentiments that are included in those documents when compared to documents using extreme, highly emotional language.

For a simple test of this idea, we applied VADER to each of the documents in the hand-labeled samples, and then split each corpus into documents with moderate sentiment (VADER scores ranging from -2 to +2, inclusive) and extreme sentiment (VADER scores either lower than -2 or higher than +2). Then we recalculated the correlation coefficient between the hand-labeled sentiment and stance values for each subsample. Table A7 shows correlations between sentiment and stance values for documents with moderate and extreme sentiment in the hand-labeled Women's March tweets corpus. Here we see that the correlation coefficient is actually a little bit higher for documents that exhibit extremely positive or negative sentiment ($z = 5.04$), suggesting that rather than the conceptual misalignment between sentiment and stance being largest at the extremes of the sentiment intensity range, it may actually be larger in the middle.

Table A7: Stance and Sentiment Intensity in the Women's March Tweets Corpus

Moderate Sentiment			Extreme Sentiment		
	Positive	Negative		Positive	Negative
Approving	8,396	2,559	Approving	4,846	1,164
Opposing	333	1,384	Opposing	161	769
		$r = 0.42$			$r = 0.48$

Table A8 shows correlations between sentiment and stance values for documents with moderate and extreme sentiment in the objectively-labeled Mood of the Nation short answer response corpus. Here we observe no statistically-discernible difference in the correlation between sentiment and stance between the moderate and extreme corpus. Likewise, Table A9 shows correlations between sentiment and stance for documents with moderate and extreme sentiment in the hand-labeled Kavanaugh tweets corpus, and shows that these correlations are essentially the same in the moderate and extreme documents here as well.

Table A8: Stance and Sentiment Intensity in the Mood of the Nation Survey Response Corpus

Moderate Sentiment			Extreme Sentiment		
	Positive	Negative		Positive	Negative
Approving	1,547	476	Approving	540	271
Opposing	777	2,511	Opposing	190	834
		$r = 0.52$			$r = 0.49$

Table A9: Stance and Sentiment Intensity in the Kavanaugh Tweets Corpus

Moderate Sentiment			Extreme Sentiment		
	Positive	Negative		Positive	Negative
Approving	262	570	Approving	259	897
Opposing	229	634	Opposing	162	647
		$r = 0.05$			$r = 0.03$

This exercise suggests that any potential misalignment between general sentiment and stance toward a specific target that we observe in a given corpus is unlikely to be a function of the intensity of the sentiment expressed in each document. The fact that an author can express an approving or opposing stance using either positive or negative language appears to apply regardless of whether that language is extremely or just moderately sentiment-laden.

References

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