Sam Cohen and Xiangming Zeng PPOL 6801

Replication Exercise #2 - Report

Introduction

Natural Language Processing (NLP) has become a much more common tool in the world of political science. One of the most common uses of NLP is sentiment analysis (i.e. how "positive" or "negative" a text is), and there are of course many applications for this in political science. However, more often than not, political scientists are more concerned with the stance of a text, not necessarily its sentiment. For this report, we will explore and attempt to replicate the findings from the research paper "Sentiment is Not Stance: Target-Aware Opinion Classification for Political Text Analysis" by Bestvater and Monroe. In this paper, Bestvater and Monroe explore the link between sentiment and stance by asking: Is it reasonable to assume that the sentiment of a document reflects the stance of that document toward the primary topic of the document?

The authors look at three primary case studies that reflect their research question well: (1) Tweets About the 2017 Women's March, (2) an Open-Ended Survey Responses About Donald Trump in his first term, and (3) Tweets About the Kavanaugh Confirmation. The authors use VADER for sentiment analysis, but also uses LLMs and human coders to predict stance, supervised learning models, and hypothesis testing. In general, Besvater and Monroe find that while there is often a correlation between ideology and support for certain political outcomes, there is little evidence to suggest that sentiment and stance exhibit the same strong relationship.

Methodology

For simplicity, we only decided to focus on the first case study in our replication. In the first case study, authors measure sentiment and stance on tweets about the 2017 Women's march, a protest march against occurring in the wake of Donald Trump's first inauguration, and one of the largest protests in United States history. The authors use a corpus of 2.5 million tweets and measure sentiment of each using VADER. These scores were further broken down further by city for geographic analysis. Authors then took a sample 20k tweets, and labeled them by hand, determining: (1) whether or not the tweet was positive or negative in sentiment, and (2) whether the tweet was pro or anti march. A BERT model was then trained on these tweets and labels to classify many of the other tweets as either pro or anti march, ro positive or negative in sentiment. Note that these values are dichotomous: 1 if pro / positive, 0 if anti / negative.

Next, the political position of each user was ranked on an ideological scale ranging from −2.5 (very liberal) to 2.5 (very conservative) using the Bayesian ideal point estimation approach, suggested and validated in Barberá (Bestvater & Monroe 2022). This approach takes into account the popular accounts each user follows.

To determine the effects of ideology on sentiment and stance, authors used logistic regression to see the relationship between ideology score and (1) VADER Sentiment, (2) BERT Stance, and (3) Human-coded stance. A SVM model to predict stance and sentiment based on the text. However, for our extension, we trained a Naive Bayes classifier to predict stance based on text instead.

Replication Results

Our code replication offered promising and similar results to those illustrated by Bestvater and Monroe. For instance, our crosstabs showing the number of positive and negative sentiment tweets and pro- and anti-march tweets showed the exact same results (Ibid). In both

cases, Positive sentiment and pro-stance tweets often went hand in hand, though negative sentiment and pro-stance tweets were the second most common cluster.

	sentiment	
stance	0	1
0	2153	494
1	3723	13242

We also recreated the logistic regression in the paper. Here, ideology scores were regressed on (1) VADER Sentiment, (2) BERT Stance, and (3) Human-coded stance. Our coefficients and results were also exactly the same as those in the paper. In general, the more liberal the twitter user, the more VADER sentiment, BERT stance, and human-coded stance tend to decrease, though the degree varies. For instance, the coefficient for ideology scores on VADER sentiment is -0.35, though this jumps to -0.76 and -1.87 for BERT stance and human-coded stance respectively (Ibid).

Logistic Regression Replication Results

	Dependent variable:		
	vader_sentiment	bert_stance	stance
	VADER Sent.	BERT Stance	Human-coded Stance
	(1)	(2)	(3)
Ideology (lib-con)	-0.353***	-0.759***	-1.868***
	(0.057)	(0.056)	(0.121)
Constant	0.837***	1.237***	2.621***
	(0.080)	(0.073)	(0.160)
Observations	928	1,501	1,501
Akaike Inf. Crit.	1,036.442	1,217.227	517.877
Note:			*p<0.05; **p<0.01; ***p<0.001

Differences

Overall, the most pressing differences resulted from the difference in original and our own VADER scores. For instance, the mean VADER score calculated for our replication analysis was 0.1565, while the mean VADER score calculated by the authors was 0.1752 (Ibid). The distribution was also slightly different when broken down by city.

There are several reasons why this might have been the case. For one, VADER is updated every few years. Scores may differ if they were calculated by the 2018 version as opposed to the 2016 version (PyPi.org 2025). In addition, the VADER sentiment scores were calculated in Python by the authors, while in contrast our analysis and calculations were done in R.

Extensions

Extension 1: Naive Bayes and Lasso classifications for stance

For Bestvater and Monroe's first case example (regarding the sentiment, stance, and ideology of tweets relating to the 2017 women's march), supervised machine learning methods

like support vector machines are employed to train models to predict the stance and sentiment based on the tweets themselves. For our extension, we created our own corpus and document frequency matrix from the tweets to see if we could recreate similar or better results with a Naive Bayes classifier or a Lasso regression classifier.

We employed a 80-20 train-test split on a corpus of 124,390 tweets. The minimum term frequency and document frequency were set as 20 and 10, respectively. For the Naive Bayes model, after multiple iterations, a smoothing parameter (alpha) of 1 was decided. An accuracy score of 86.3 percent was observed, with a recall of 96.4 percent and a precision score of 86.5 percent. The F1 score was 91.2 percent. Our model also even performed better than Bestvater and Monroe's SVM stance model, which exhibited a F1 score of 81.7.

The results of the Lasso regression model are even more impressive, with an accuracy of 90.9% and an F1 score of 91.2%. In addition to its strong predictive performance, the Lasso model also offers valuable insights into the sentiment expressed by different stance groups. After training the model, we identified the most predictive features for each stance by selecting those with the largest positive and negative coefficients. As shown in Figure 1, the features on the left are the strongest predictors of the pro-march stance, since the label "1" corresponds to pro-march. Most of these features are hashtags or geographic references, which are relatively neutral in tone. In contrast, the features most predictive of the anti-march stance are highly emotional, including numerous conservative phrases, profanities, and negative expressions. This finding supports the paper's central argument: sentiment is not a perfect proxy for political stance. If it were, we would expect many strongly positive terms among the top predictors of the pro-march group—but that is not what we observe.

coefficient <dbl>	feature <chr>	coefficient <dbl>	feature <chr>
5.389272	#strength	-5.811211	#tcot
3.224309	#polit	-5.012559	pin-up
3.034790	sen	-4.772593	illeg
2.901997	reno	-4.676034	#prolif
2.770356	#womensmarchjxn	-4.603949	whore
2.622214	#womansmarch2017	-4.364715	pro-lif
2.512320	ohio	-4.346161	disgrac
2.468688	burst	-4.292443	maga
2.442510	sc	-4.270926	isi
2.397230	#portland	-4.213256	hypocrisi

Figure 1. Most predictive features

Extension 2: ALC word embeddings

We applied ALC embeddings to the sample of 20k tweets labeled by hand. The pre-trained fastText transformation matrix and embeddings provided by the ALC embeddings

website were used. Generally, it works well in identifying the nearest neighbors of two word patterns (march-related terms & trump-related terms) by stance.

Why did we select “march” and “trump”? It’s clear for “march” since we were studying tweets of a march. But why “trump”? The reason is this protest was triggered by Trump’s policy positions, so finding the nearest neighbors for trump-related terms might provide valuable insights.

The second question is, why did we use word patterns, rather than using single words directly? The reason is that since there is no stemming in the preprocessing of ALC embeddings training, we didn’t apply stemming too in order to keep the pre-processing close to what ALC embedding team did for training. Therefore, if we only rely on single words, we risk omitting lots of relevant concepts, especially in dealing with tweets containing so many usernames and hashtags.

We can tell from Fig 2. that sentiment of march-related terms differs significantly across stance groups. The closest neighbors for the anti-march group are mostly profanities and negative expressions, while the closest neighbors for the pro-march group are emotionally neutral (like “yesterday”) or positive (like “proud” and “thank”), even with some expressions of sense of belonging (like “folks” and “we’re”). Moreover, there are only two features shared by the two groups. Therefore, if we only have tweets talking about the march, sentiment may be a better proxy of political stance.

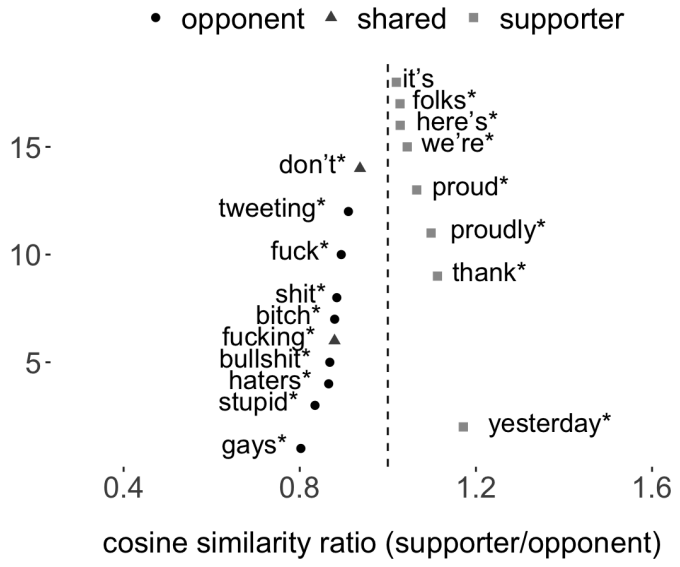

Figure 2. Closest neighbors for march-related terms

However, as shown in Fig 3, when people are talking about Trump, sentiments are relatively consistent in two stance groups. There are 8 out of 10 closest neighbors shared by the two stance groups. Moreover, 7 out of 8 shared features are profanities or negative expressions (if we treat “don’t” and “anti-” as negative). This finding also resonates with the main argument of the paper. When pro-march individuals are talking about Trump, they are negative too.

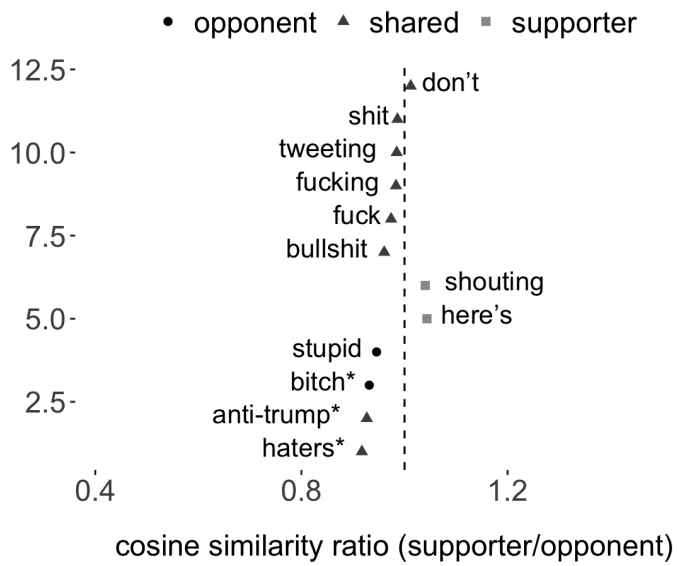

Figure 3. Closest neighbors for Trump-related terms

Extension 3: Wordscore

Inspired by Mitra (2021), we tried to use the hand-coded stance labels (0 = anti-march, 1 = pro-march) to conduct wordscore. After playing around with the hyperparameter, we find when smooth equals 0.01, features with the largest positive and negative scores are most interpretable.

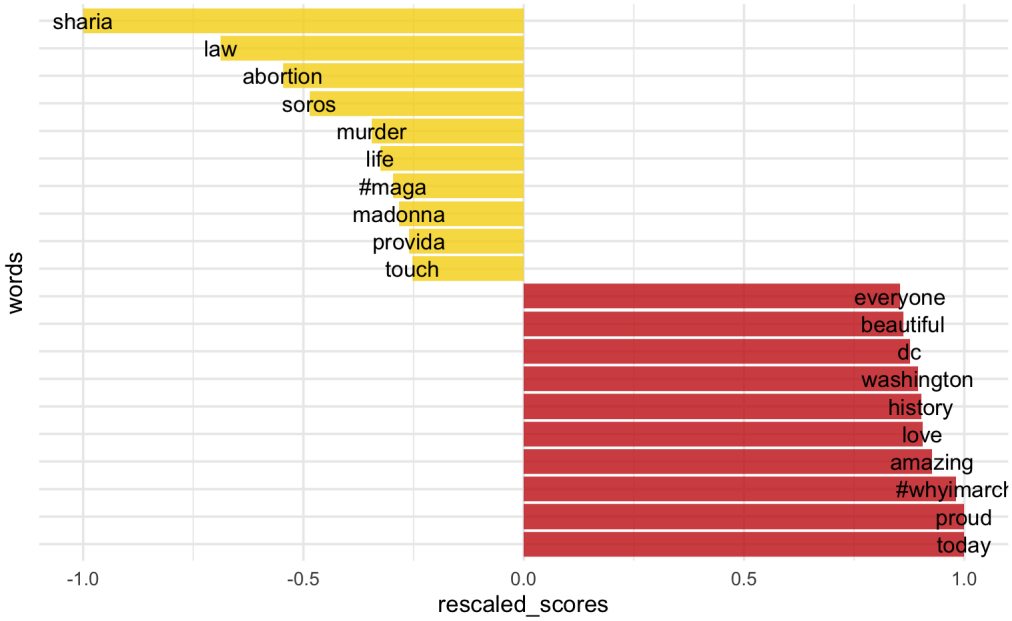

Figure 4. Features with the largest positive and negative rescaled scores

As shown in Fig 4, there are lots of positive expressions in the positive end, like “beautiful,” “love,” “amazing,” and “proud.” However, we see few negative expressions in the negative end (if we treat “murder” as negative). In other words, they are not predictive enough for the anti-march stance. This result aligns with the ALC embedding analysis of Trump-related terms, which shows that profanities and negative expressions are common across both stance groups when discussing certain topics, resulting in lower absolute word scores for those terms.

Autopsy

The most prominent differences resulted from the difference in original and our own VADER scores. This may be due to (1) package version differences, and (2) different softwares (R vs. Python).

In addition, since our corpus (tweets) does not particularly resemble Wikipedia, we tried to train the local transformation matrix and GloVe embeddings. However, the nearest neighbors identified based on maximum cosine similarity to march-related terms did not appear particularly insightful, as shown in Fig 5. We consulted Wirsching et al. (2025), who suggest: “If their corpus is too small to fit local models, we recommend using our estimated A matrix, and carefully checking its validity.” We believe that the limited size of our corpus explains the unsatisfying performance of both the A matrix and the embeddings we trained.

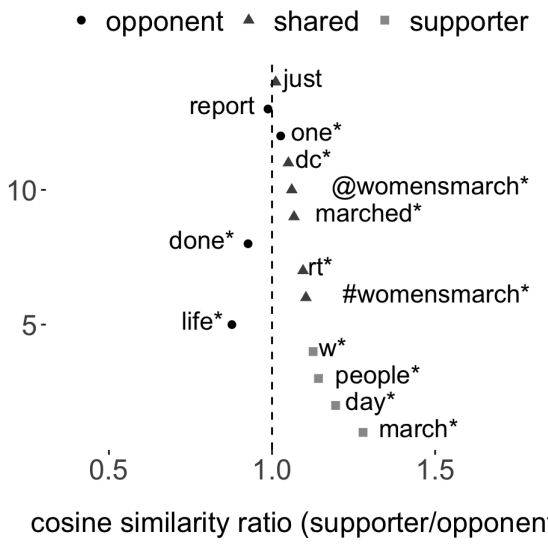

Figure 5. Closest neighbors for march-related terms with A matrix and embeddings trained by us

References

Bestvater, S. E., & Monroe, B. L. (2023). Sentiment is Not Stance: Target-Aware Opinion Classification for Political Text Analysis. Political Analysis, 31(2), 235–256. doi:10.1017/pan.2022.10 https://www.cambridge.org/core/journals/political-analysis/article/sentiment-is-not-stance-targetaware-opinion-classification-for-political-text-analysis/743A9DD62DF3F2F448E199BDD1C37C8D

Mitra, Alessio, Do Speeches at the UN General Assembly Affect International Aid Allocation? (March 13, 2021). Available at SSRN: https://ssrn.com/abstract=4034353 or http://dx.doi.org/10.2139/ssrn.4034353

PyPi. (2025). vaderSentiment 3.3.2. pypi.org. (Accessed 2025). https://pypi.org/project/vaderSentiment/#history

Wirsching, E. M., Rodriguez, P. L., Spirling, A., & Stewart, B. M. (2025). Multilanguage Word Embeddings for Social Scientists: Estimation, Inference, and Validation Resources for 157 Languages. Political Analysis, 33(2), 156–163. https://doi.org/10.1017/pan.2024.17