On the Reliability and Validity of Detecting Approval of Political Actors in Tweets

Indira Sen
GESIS
Indira.Sen@gesis.org

Fabian Flöck
GESIS
Fabian.Floeck@gesis.org

Claudia Wagner
GESIS & RWTH, Aachen
Claudia.Wagner@gesis.org

Abstract

Social media sites like Twitter possess the potential to complement surveys that measure political opinions and, more specifically, political actors' approval. However, new challenges related to the reliability and validity of social-media-based estimates arise. Various sentiment analysis and stance detection methods have been developed and used in previous research to measure users' political opinions based on their content on social media. In this work, we attempt to gauge the efficacy of untargeted sentiment, targeted sentiment, and stance detection methods in labeling various political actors' approval by benchmarking them across several datasets. We also contrast the performance of these pretrained methods that can be used in an off-the-shelf (OTS) manner against a set of models trained on minimal custom data. We find that OTS methods have low generalizability on unseen and familiar targets, while low-resource custom models are more robust. Our work sheds light on the strengths and limitations of existing methods proposed for understanding politicians' approval from tweets.

1 Introduction

Measuring public opinion accurately and without systematic errors is as vital for a functioning democracy as it is for scholars to understand society. Survey methodologists have developed techniques over several decades to precisely quantify public opinion. The American Association for Public Opinion Research (AAPOR) stated in their recent task force report that public opinion research is entering a new era, where digital traces would play an important role (Murphy et al., 2014). Increasingly, since the first steps were made by O'Connor et al. 2010, numerous studies have assessed the efficacy of such traces, especially social media, in measuring public opinion as a complement to polls. The run for social media approaches is not surprising, as they promise a continuous public opinion estimate based on millions of data points.

Tweet	Untargeted Sentiment	Targeted Sentiment	Stance	Approval
Trump is the only candidate I fully support	positive	positive	favor	approval
What makes me angry is the lying media brazenly attacking President Trump.	negative	positive	favor	approval
Jeb Bush is the only sane candidate in the republican lineup	positive	none	against	disapproval

Table 1: Different types of NLP measurements that can be used to understand a tweet's **approval of a predefined target** (here, Donald Trump): Untargeted/overall sentiment (UTS), targeted sentiment (TS) and stance (ST). UTS can easily fail to measure approval of the target if several potential targets are mentioned or the actual target is not explicitly present. TS cannot measure indirect opinions where the target is not mentioned, whereas ST methods are designed for this task as well.

Social-media-based metrics require new approaches, which bring forth new challenges (Olteanu et al., 2016). Sen et al. 2019 describe two primary sources of errors: *representation errors*, due to how results are inferred for the target population and *measurement errors*, due to how the target construct is measured. Researchers have made substantial advances in understanding and adjusting for representation errors (Pasek et al., 2019; Barberá, 2016). Yet, there is still a gap in knowledge about whether the lack of effectiveness of social media-based estimates is *also* due to measurement errors, i.e., the operationalization of the target construct – approval. While previous research has used external data such as polling results to (in)validate the efficacy of automated *aggregate* approval measures from social media, the *fine-grained (mis)measurement of approval on a post level* has yet to be studied.

The building blocks for measuring approval with

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 1413–1426,
November 16–20, 2020. ©2020 Association for Computational Linguistics

social media are usually the textual utterances by users.[1] Related work predominantly focuses on the largest publicly available social media platform, Twitter, and employs methods ranging from sentiment lexicons (O'Connor et al., 2010; Pasek et al., 2019) to machine learning approaches (Marchetti-Bowick and Chambers, 2012; Barberá, 2016) for analyzing approval in individual tweets. Several natural language processing (NLP) approaches have been proposed to, or can be amended to, measure approval. They can be segmented into three broad classes: untargeted sentiment detection, targeted sentiment detection, and stance detection (see Table 1).[2]

Untargeted sentiment is a popular choice for measuring approval (Pasek et al., 2019; O'Connor et al., 2010; Pasek et al., 2018), possibly due to the availability of several methods that can be used without much overhead in an off-the-shelf (OTS) manner. Yet, cognitive scientists contend that attitudes such as approval are tied to an object of approval (Bergman, 1998), and untargeted sentiment, in comparison to targeted sentiment and stance, might not be the best proxy for it (c.f Table 1). Indeed, as it is the most sophisticated family of methods and aligned with what we term "approval", **stance detection** by design typically outperforms sentiment detection methods within shared tasks (e.g., SemEval) aiming to measure targets' approval. While stance may indeed be a more robust theoretical proxy, a potential obstacle towards using stance detection, instead of untargeted sentiment analysis, is the lack of OTS methods available. Even for methods that do exist, the developers intentionally or unintentionally tune their methods towards benchmark datasets (e.g., by exploiting the fact that a dataset is collected based on particular hashtags). It is thus likely that complex methods are tuned to linguistic markers of benchmark datasets and only perform well on those or similar datasets (Linzen, 2020). In this light, it is unclear if such methods can be used "tout court" on novel datasets and targets.

Therefore, we investigate the following use case: Measuring how respondents or users feel towards a certain topic or entity (which we call *target*), such as the president (O'Connor et al., 2010; Pasek et al., 2019) or presidential candidates (Barberá, 2016), where the outcome is captured on a continuum between approval and disapproval or some equivalent, e.g., favor, neutral, against. While different terms like "viewpoint", "support" or "stance" can be ascribed to this measurement, we will henceforth call it "*approval*"; this mirrors the long-standing measurement tradition in survey research to ask for the approval of political actors and issues, usually also indicated on a scale with synonymous extremes.[3] We investigate the design choices to be made by a researcher to increase reliability and validity of the measurement. [4]

Our Contributions. To investigate how well automated methods capture approval on a fine-grained tweet level, we systematically compare the validity and reliability of (i) "off-the-shelf" (OTS) usage of methods that require minimal effort to (re)use, and (ii) customized low-resource methods, leveraging popular supervised text classification models,[5] trained on varying, small-scale quantities of in-domain data, to simulate a scenario where individual datasets are labeled with realistically expendable effort (Adams-Cohen, 2020; Hughes et al., 2020). Across five different datasets, spanning seven targets, we benchmark the performance of twelve methods: eight OTS methods that have been used in the past for assessing approval or are exemplary for different types of NLP approaches that have been proposed for understanding concepts akin to approval, and four customized low-resource methods. We find more complex supervised OTS methods, especially targeted methods, do not generalize well to unseen targets, i.e., targets that are not present in the training data of these methods. But they also have high variation on familiar targets, where they struggle with measuring instances of indirect stance and absence of stances. Low resource custom methods outperform OTS methods for both types of targets. Our systematic analysis identifies and highlights gaps in current

[1] These can further be aggregated per user (Cohen and Ruths, 2013), but we focus on the much more common practice of measuring post-level opinions.

[2] Stance detection here is different from rumorstance detection (Kochkina et al., 2017) and argument stance detection (Lippi and Torroni, 2016), where the task is to infer the speaker's reaction to a potential rumor or argument, respectively.

[3] For example, Gallup's poll on presidential approval has remained virtually unchanged for decades (McAvoy, 2008).

[4] Quinn et al.: "The evaluation of any measurement is generally based on its reliability (can it be repeated?) and validity (is it right?)." In this work, by validity, we refer to external validity or generalizability, while reliability refers to repeating the same measurements under different conditions.

[5] In this work, we differentiate between *models* which are machine learning models that can learn from data, and *methods* which have already been trained and can be re-used without further training or fine-tuning.

methods for the measurement of approval and implies that **even though targeted sentiment and stance are better proxies for approval than untargeted sentiment, current targeted methods cannot be used in an OTS manner for measuring approval.** Our code is available at `https://github.com/gesiscss/political_approval`

2 Methods for Measuring Approval

In the section, we describe widely used methods that have been applied to mine public opinion on Twitter, particularly approval of political actors. Evaluating all pertinent methods and their varying implementations is beyond the scope of this work, therefore we choose popular approaches or those whose implementations are widely available.

We describe the three above-mentioned categories of approaches (summarized in Table 1) which can be used as proxy measures for approval or disapproval of targets.

2.1 Untargeted Sentiment

Untargeted sentiment refers to the *overall* sentiment of a sentence or document, regardless of targets mentioned. Prominent and easy-to-use representatives of untargeted sentiment methods are lexicons of positive and negative words. The word lists are hand-curated and are usually not adapted to each target dataset they are applied to. They are typically used to annotate words in documents and the ratio of positive to negative words in a document may function as an indicator of opinion (O'Connor et al., 2010). To arrive at a measurement of approval, the document for which overall sentiment is calculated is either assumed to be about the target *a priori* via the collection process of the corpus (O'Connor et al., 2010; Pasek et al., 2018), or is labeled as such through heuristics or named entity recognition. In this work, we compare various lexicons which have been used in past public opinion analysis literature: **MPQA** (Hu and Liu, 2004) and **LabMT** (Dodds et al., 2011) used by O'Connor et al. and Cody et al., respectively to understand approval of President Obama. **VADER** (Hutto and Gilbert, 2014), which is a lexicon combined with a heuristic-based preprocessing engine for understanding syntactic characteristics of sentences such as negation, was recently used to understand stance towards the economy (Conrad et al., 2019).

In contrast to lexicons, we also explore fully supervised methods including **SentiStrength (STS)** (Thelwall, 2017), a widely-used lexicon-based supervised method[6] and **SentiTreeBank (STB)** (Socher et al., 2013), trained on human-annotated web content such as online reviews. While both STS and STB include syntactic dependencies so they can account for negations and modifiers, they are target-independent and can therefore capture the overall sentiment of a tweet rather than sentiment towards a particular entity.

2.2 Targeted Sentiment

The task of Targeted Sentiment Analysis (TS) is, given a sentence, to infer the sentiment of the author towards a predefined topic or entity.[7] **TD-LSTM** (Tang et al., 2016) is a Recurrent Neural Network based approach that also takes into account syntactic dependencies, trained and tested on a Twitter dataset with tweets towards various entities and topics like Bill Gates, Lady Gaga, and Donald Trump, annotated by crowdworkers (Dong et al., 2014). TD-LSTM achieved state-of-the-art performance (69% Macro F1) on the aforementioned Twitter targeted sentiment dataset. To translate targeted sentiment to stance or approval, a function is commonly defined that transforms negative sentiment scores to disapproval or "against" and positive sentiment to approval or "for", with a residual category of "neutral" for mid-range or inconclusive scores.

2.3 Stance

Stance detection refers to a set of loosely connected tasks in NLP such as argumentation mining and rumor verification.[8] In this work, we focus on the specific case of stance detection, closely related to TS, which is the task of inferring whether a document is written in favor or against the given target. Stance detection and TS differ in that the author may take an *indirect* stance without explicitly mentioning the target. While various stance detection methods exist, we focus on two prominent example methods that have been developed specifically for detecting stance on Twitter. Mohammad et al. introduce a strong Linear SVM-based method (**SVM-SD**) trained on a stance-annotated tweet corpus with

[6] Sentistrength, for example, has been used to assess the sentiment of tweets mention German politicians: https://data.gesis.org/tweetskb/

[7] The task is closely related to, but distinct from, Aspect-Based Sentiment Analysis. More specifically, TS is described as Targeted Non-aspect-based Sentiment Analysis (TN-ABSA) where "the object of the analysis is simply the target entity." (Pei et al., 2019).

[8] See (Küçük and Can, 2020) for a comprehensive survey on various types of stance detection tasks.

Method	Supervised/ Unsupervised	Type	Output	Reference	Implementation
VADER	unsupervised	UTS	compound score between [-1, 1]	Hutto and Gilbert	https://github.com/cjhutto/vaderSentiment
MPQA	unsupervised	UTS	Ratio of positive and negative score	Hu and Liu	https://mpqa.cs.pitt.edu/lexicons/subj_lexicon/
LabMT	unsupervised	UTS	Ratio of positive and negative score	Dodds et al.	https://hedonometer.org/words/labMT-en-v1/
Sentistrength (STS)	supervised	UTS	[-5,5]	Thelwall	http://sentistrength.wlv.ac.uk/
SentiTreeBank (STB)	supervised	UTS	[very positive positive, neutral, negative, very negative]	Socher et al.	https://nlp.stanford.edu/sentiment/treebank.html
TD-LSTM	supervised	TS	[negative, none, positive]	Tang et al.	https://github.com/jimmyyfeng/TD-LSTM
SVM-SD	supervised	ST	[favor, none, against]	Mohammad et al.	
DSSD	supervised	ST	[favor, none, against]	Augenstein et al.	https://github.com/sheffieldnlp/stance-conditional
Custom (LR)	supervised	ST	[favor, none, against]		
Custom (SVM)	supervised	ST	[favor, none, against]		
Custom (MNB)	supervised	ST	[favor, none, against]		
Custom (BERT)	supervised	ST	[favor, none, against]		

Table 2: Overview of the tweet-level methods used to understand approval. The first eight are **off-the-shelf**, i.e., not trained on any novel data while the bottom four are **custom**, i.e., trained on minimal in-domain data. We categorise methods based on their training procedure (supervised or unsupervised), the type of proxy they measure, untargeted sentiment (UTS), targeted sentiment (TS) or stance (ST), and describe their output. Since the custom methods are trained on data annotated for stance, we also consider them to be of that type. We also include the source of implementation of off-the-shelf methods when available.

character and word n-grams that outperformed all submissions in the SemEval 2016 Stance Detection shared task A (Mohammad et al., 2016).

Secondly, for their **Distant Supervised Stance Detection (DSSD)** method, Augenstein et al. train an LSTM on tweets where stance towards various entities or topics is labeled (cf. the SemEval 2016 Stance Detection shared task A dataset (Mohammad et al., 2017)). However, the final goal of this method is to label stance in tweets towards Donald Trump, which was not included as a potential target in the training data (shared task B). To improve prediction performance for an unknown entity (Trump in this case), the authors leverage a large collection of tweets containing keywords relevant to Trump, weakly labeled based on the presence of certain keywords or hashtags such as 'MAGA' and '#yourefired', in conjunction with a bidirectional LSTM.[9] We include this method since it achieved high performance (average of 59% macro F1 on favor and against classes) on the shared task.

3 Use case scenarios

We now describe the two scenarios we explore as realistic options faced by a CSS researcher aiming to measure approval towards political actors on Twitter with their own dataset and/or targets.

3.1 "Off-the-shelf" usage

As our first scenario, we assume that a researcher does not have the resources to label their novel data and/or retrain their own model on this data and targets they are working with. A low-threshold solution is (i) the usage of dictionary-based methods or (ii) the use of existing supervised methods pretrained on a different corpus and potentially different targets.

As dictionaries are not trained by design, we employ them with only minor adaptions to their preprocessing pipelines. Due to the lack of a standardized processing pipeline for LabMT and MPQA, and to maintain consistency within the lexicons, all three of them are used in conjunction with VADER's preprocessing engine. MPQA and LabMT which yield ratio of positive and negative scores are converted to three classes reliant on a value greater (favor), lesser (against) or equal (none) to zero. Following past literature (Hutto and Gilbert, 2014), we use -0.1 and 0.1 as the threshold for converting VADER scores to positive (favor) and negative (against), respectively. STS and STB are used with their pretrained models. We reimplement TD-LSTM using the code made available by the authors (c.f Table 2 and Appendix C). TD-LSTM and STS provide scores of positive, negative and none which can be mapped to the aforementioned stance classes. For STB, we collapse the five-class output to three-class, by combining very negative (very positive) and negative (positive). Like TD-LSTM, we re-implement DSSD, and replicate SVM-SD based on Mohammad et al. 2017.

[9] Generating weak labels may require domain knowledge and is not equally plausible for all targets, especially for novel targets.

Target	Dataset	against		favor		none		Total
		direct	indirect	direct	indirect	direct	indirect	
Trump	CONS	156	62	53	20	9	3	303
	MTSD	620	0	989	0	454	0	2063
	PRES	387	1	144	0	96	0	628
	SEB	165	134	146	2	6	254	707
Macron	PRES	234	0	135	0	177	0	546
Clinton	CONS	78	19	109	46	3	5	260
	MTSD	507	0	220	0	262	0	989
	SEA	107	64	42	3	2	76	294
Zuma	PRES	363	3	134	0	122	0	622
Widodo	PRES	101	0	150	0	168	0	419
Erdoğan	PRES	378	1	81	0	141	0	601
Putin	PRES	416	0	103	0	99	0	618

Table 3: **Datasets.** The datasets used for evaluating all methods, related to different political actors and approval (stance) distribution. We use a held-out sample of this data, stratified on stance, to train low-resource custom methods on minimal data (195 tweets from each target) and use the rest for testing the OTS and custom methods.

3.2 Customized Training

For this scenario, we assume that limited resources are available to label the dataset to be analyzed towards the desired target, and that commonly available NLP models, particularly those that have been used for text classification, can be employed to train custom methods accordingly. Training data for novel targets can be expensive to generate, so we train models on a held-out minimal proportion of the test datasets (Table 3) to obtain target-specific stance methods, similar to Mohammad et al. (2017), but on a fraction of the data; 195 datapoints from each target.[10] We decide on this threshold based on the least amount of labeled data required to outperform the best performing OTS method, as further explained in Appendix A.

We consider a small number of concrete manual labels of tweets as the most realistic scenario. We do not consider using weak labels "low effort", since (i) they have to be carefully selected for each target, e.g., by a domain expert and be sufficiently tailored to the target, such as a politician-specific hashtag, and (ii) a large amount of labels would be required for retraining a method such as DSSD, which is not feasible for each dataset used in our evaluation, nor in practice in many cases.

For the custom models, we also remove stopwords (except 'not') and use unigram features to train four different types of models that are popular for text classification tasks: Logistic Regression (LR), Multinomial Naive Bayes (MNB), a Support Vector Machine (SVM) and finetuned BERT (Devlin et al., 2019). For LR, MNB and SVM, we perform five-fold cross-validation and grid search to tune hyperparameters. For BERT, 10% of the dataset is used as a validation set (c.f Appendix C for hyperparameter configurations). Our objective is not to build a state-of-the-art classifier with optimal performance, but to understand how methods utilizing minimal training data compare against OTS methods.

3.3 Baselines

To emulate an "absolute minimal effort" scenario we set up three baselines. The first is a **random baseline**, a classifier that randomly assigns a stance label (either favor, against or none) to each tweet. The second and third baseline are based on a classifier that labels every instance with the majority label for the dataset (independent of targets) (**majority-dataset**) or target (**majority-target**).

4 Experiments

Previous research established the validity of social media measures through correlations with external data sources like polls and surveys (O'Connor et al., 2010; Pasek et al., 2018; Barberá, 2016). We argue that this entangles different types of errors, such as the lack of demographic match between polls and social media users and the effect of the platform's affordances on textual expressions. By focusing on a controlled dataset of human-annotated approval at a tweet level, we can rule out confounding factors to a higher degree. Furthermore, as we see in Table 1, stance is a better proxy for approval than targeted and untargeted sentiment. Therefore, we compare the performance of the previously described methods over five different datasets that form the gold standard of stance ($\sim$ approval). Using datasets spanning different targets as well as different time periods helps us gauge the generalizability and robustness of methods. In this section, we describe our experimental setup and datasets used for evaluation and custom training.

4.1 Experimental Setup

Evaluation parameters. We use **Macro-F1** (which weights all classes equally) across all three classes to analyze performance. To assess cross-dataset and cross-target performance, we compute the mean, standard deviation and the upper (high) and lower (low) bounds of 95% confidence interval. To account for possible variance, all methods are evaluated based on average performance on the

[10] Since, different targets have varying amount of data, 195 tweets constitutes 5.5% of the Trump data and 12%-46% of the other targets.

evaluation datasets (Table 3) over 5 runs.

4.2 Datasets

We evaluate OTS and custom methods on the following datasets. While some of these datasets have common targets, for example, Trump is present in four of them, they are all collected in different periods of time, with different keywords (c.f Appendix B). All datasets have stance labels of 'favor', 'against', and 'none' towards the targets.

SemEval A and B. The SemEval-2016 task 6 dataset (Mohammad et al., 2017) contains topic-tweet pairs, on controversial subjects. Since our analysis is restricted to political actors, we use the portion of the task A test dataset with stance towards Hillary Clinton (**SEA**) and the task B dataset with stance towards Donald Trump (**SEB**).

Constance (CONS). Joseph et al. (2017) released a dataset containing stance towards Trump and Clinton. The authors use this dataset to understand how different annotation contexts affect crowdworkers' performance in labeling tweets for stance. The authors annotate tweets based on various contextual information such as the profile details of the tweet author.

MTSD. Sobhani et al. (2017) released a dataset where each tweet has stance towards more than one target (multi-target stance detection). The authors collected data about four presidential candidates of the US 2016 elections using related hashtags, selecting three target pairs: Donald Trump and Hillary Clinton, Donald Trump and Ted Cruz, Hillary Clinton and Bernie Sanders. We only include those tweets where one of the targets is either Trump or Clinton.[11]

Presidents (PRES). van den Berg et al. (2019) collect a dataset of tweets mentioning presidents of six G20 countries by various naming forms, which are annotated for stance. The authors investigate the role of naming variation in stance towards presidents. To do so, the authors collect tweets three query types: last-name, #first-name and first-name + (last-name/country). They then leverage crowdworkers for annotating the stance in these tweets.

4.3 Experimental Design

We run the following two experiments to assess validity and reliability respectively.

Experiment 1. We evaluate performance of methods across all targets. This allows us to assess the external validity of various OTS methods by measuring how well they generalize to unfamiliar targets (OTS scenario) compared to custom methods that have seen a minimal portion of the data related to such targets (custom training scenario).

Experiment 2A. We evaluate the performance of methods for the target Donald Trump, a target familiar to some OTS methods like TD-LSTM and DSSD, across multiple datasets (CONS, MTSD, PRES and SEB). This allow us to assess the reliability of methods in measuring the same construct ('approval of Trump'), across multiple settings which span over different time periods and employ different data collection strategies.

Experiment 2B. The advantage of stance over TS is indirect stances.[12] Therefore, we also investigate how well various methods perform on indirect stance. Here, direct stance refers to when the target is mentioned by name. For example, tweets with indirect stances towards Trump mention neither his firstname, lastname nor his Twitter handle (@realdonaldtrump). They may refer to him indirectly, say, via epithets ('@potus') or his association to other subjects or entities (example 3 in Table 1).

5 Results

We now describe our findings from the the experiments described in the previous subsection. We compare the performance of methods across different targets in Table 4 and across datasets that have been collected in different ways but include one target (Trump) in Table 5. Finally, we investigate the performance on indirect and absence of stance.

5.1 External Validity: Performance across Targets

To compare the external validity (which refers to the generalizability) of various methods we present their performance across different targets in Table 4. First, the low performance of baselines demonstrate that inferring stance is a hard task. Sentiment lexicons (VADER, LabMT, MPQA) perform surprisingly well and outperform more complex OTS methods. This indicates that targeted OTS methods do not work well for unfamiliar targets which they have not seen during the learning phase. Therefore, it is not advisable to use such methods for novel targets. **Our results indicate that targeted OTS methods, like DSSD and TD-LSTM, should not**

[11] For our purpose, we only use the stance towards either of these two as our final stance label.

[12] These include references to the target "through pronouns, epithets, honorifics, and relationships."(Mohammad et al., 2017)

Method	Targets							Mean F1	Std	95% CI	
	Clinton	Erdoğan	Macron	Putin	Trump	Widodo	Zuma			high	low
majority-target	22.30	25.79	19.99	26.85	19.45	19.08	24.69	22.59	2.97	28.42	16.77
majority-dataset	23.92	25.79	19.99	26.85	38.44	13.00	24.69	24.67	7.11	38.61	10.73
SVM-SD (ST)	28.08	30.23	29.37	33.62	23.54	16.30	30.41	27.37	5.32	37.79	16.94
STB (UTS)	28.19	30.64	32.53	37.83	23.60	21.23	31.08	29.30	5.17	39.43	19.17
random	31.57	31.65	29.09	28.45	33.16	32.31	30.65	30.98	1.58	34.08	27.89
STS (UTS)	31.23	30.52	29.23	29.84	30.93	33.74	32.01	31.07	1.38	33.77	28.37
DSSD (ST)	35.08	25.29	28.98	33.06	44.12	25.02	34.59	32.30	6.17	44.40	20.21
BERT (custom)	26.45	35.42	33.86	27.06	40.27	41.23	25.19	32.78	6.17	44.87	20.70
TD-LSTM (TS)	19.74	33.13	40.10	37.51	34.47	47.77	39.48	36.03	7.97	51.65	20.41
MPQA (UTS)	31.34	41.69	34.24	33.56	32.98	47.00	37.50	36.90	5.21	47.11	26.69
LabMT (UTS)	34.19	37.73	42.12	36.44	33.33	48.10	38.37	38.61	4.71	47.85	29.38
SVM (custom)	33.33	45.45	41.61	40.07	43.96	51.57	34.03	41.43	5.95	53.09	29.77
VADER (UTS)	36.72	44.81	**43.49**	43.51	36.97	50.32	44.24	42.86	4.38	51.45	34.27
MNB (custom)	40.23	47.22	41.60	43.36	40.64	50.63	42.81	43.79	3.53	50.70	36.87
LR (custom)	**44.08**	**52.11**	42.51	**48.49**	**47.23**	**53.18**	**46.40**	**47.71**	3.63	54.82	40.60

Table 4: **Overview of the performance (Macro F1) of different methods across all targets.** UTS = Untargeted Sentiment, TS = Targeted Sentiment, ST = Stance. Cross-target performance for supervised off-the-shelf methods are poor with high variability. Unsurprisingly, lexicons have more stable performance, but surprisingly, outperform targeted methods. The LR custom method, trained on minimal data, performs best. The results indicate that targeted off-the-shelf methods, like DSSD and TD-LSTM, are not 'general-purpose' since their performance is as good as or even worse than untargeted lexicons like VADER and MPQA for targets that are new to them.

be considered 'general-purpose', and that their performance is as good or even worse than untargeted lexicons like VADER, LabMT and MPQA for new and unseen targets such as Macron and Putin.[13] The LR custom method performs best for all targets except Macron (where the best method is VADER), while BERT performs poorly, possibly due to insufficient training data, indicating that a simple, high-bias classifier performs better than complex methods, OTS and custom alike, if the amount of available training data is low.

5.2 Reliability: Performance across Datasets for Donald Trump

Since it is not surprising that targeted methods have low generalisation to unseen targets, we now evaluate them on a familiar target: Trump (Table 5). DSSD was trained on weak Trump labels, while the training data for TD-LSTM also contained tweets with sentiment towards Trump. When comparing the performance of different methods across different datasets with approval towards Trump, we find that targeted methods perform far better than they had for unseen targets but still show a wide range of variation. DSSD for example, which achieves strong results on SEB, drops in performance across all other datasets. The inconsistency and reduced performance of supervised methods could be due to difference in label distribution in train and test sets (dataset drift), and the fact that scientists often finetune their methods for a specific tasks which may lower the generalizability. In this case, the heuristics used to generate weakly labeled data to train DSSD may not hold in different time periods. One can finetune these methods for each dataset separately, but this is not always feasible due to the lack of computational skills and/or availability of data; either weakly labeled data or larger quantities of 'strongly' labeled data required for training deep learning models. Our results also indicate that even if weak labels are generated for a specific target, they might not help the method trained on them to generalize beyond the dataset from which weak labels were generated.

MTSD is the most difficult dataset to classify for most methods, possibly due to the presence of multiple entities and different stances towards them. TD-LSTM outperforms stance detection methods on all non-SEB datasets but performs poorly on SEB. As seen for other targets, the LR custom method surpasses OTS methods in mean F1, while BERT performs poorly. **Our results indicate that**

[13] To rule out issues due to model architecture, we also finetune BERT models on weak labels used to train DSSD. This model slightly outperforms DSSD but still has worse performance than VADER and the LR custom method.

low resource methods might be more advantageous than OTS method, when the sample that needs to be analyzed may have different characteristics (say, time period or keywords used for tweet selection) to the OTS methods' training data, even if the target entity is the same.

Method	Trump Datasets				Mean F1	Std	95% CI	
	CONS	MTSD	PRES	SEB			high	low
majority-target	27.86	15.39	25.44	19.90	22.15	4.85	31.66	12.63
majority-dataset	27.86	21.65	25.44	19.90	23.71	3.12	29.83	17.59
SVM-SD (ST)	26.40	15.54	29.24	28.96	25.03	5.59	36.00	14.07
STB (UTS)	29.98	18.18	30.96	28.20	26.83	5.09	36.81	16.86
random	25.93	32.85	30.01	32.15	30.24	2.70	35.52	24.95
MPQA (UTS)	23.59	33.26	28.06	37.64	30.64	5.30	41.02	20.25
STS (UTS)	32.50	30.58	32.09	27.88	30.76	1.81	34.31	27.21
BERT (custom)	31.25	40.66	30.39	22.22	31.13	6.53	43.93	18.32
LabMT (UTS)	29.26	31.37	36.12	35.86	33.15	2.94	38.91	27.40
TD-LSTM (TS)	32.44	30.98	39.87	38.84	35.53	3.87	43.13	27.94
VADER (UTS)	29.52	35.65	**40.11**	37.96	35.81	3.96	43.57	28.05
SVM (custom)	32.05	41.88	31.93	41.05	36.73	4.75	46.04	27.42
MNB (custom)	**34.25**	38.84	34.71	39.57	36.84	2.38	41.51	32.17
DSSD (ST)	31.76	29.72	34.17	**60.59**	39.06	12.53	63.61	14.50
LR (custom)	31.66	**43.88**	33.85	47.30	**39.17**	6.58	52.07	26.28

Table 5: **Overview of the performance of methods measuring support of Donald Trump** Targeted supervised methods like TD-LSTM and DSSD outperform sentiment lexicons, with a few exceptions such as VADER (comparable performance). DSSD performs notably worse on datasets other than SEB and shows high standard deviation. Custom LR methods outperform off-the-shelf methods, even for familiar targets.

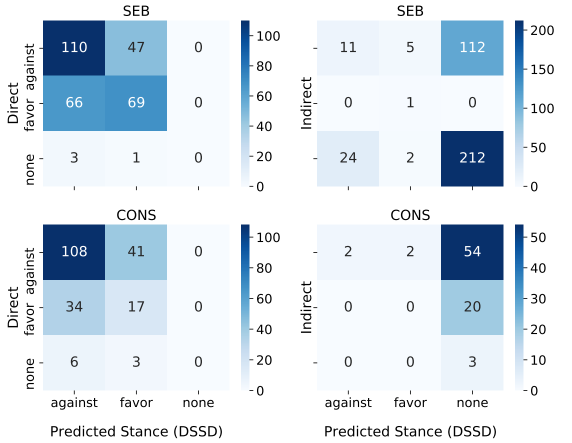

Figure 1: Confusion matrices of DSSD on SEB and CONS disaggregated by directness of stance. DSSD labels most of the indirect cases in SEB as 'none' and has difficulties assessing indirect stance in CONS.

5.3 Error Analysis

We analyze why current OTS methods fail by taking a closer look at their performance on two dimensions: directness and presence of stance. As a case study, we focus on DSSD and compare its performance on SEB (F1 score of 60.6%), to other datasets, where performance is relatively lower.

Tweet	Dataset	Stance Type	True Stance	Predicted Stance
After today SCOTUS has passed more legislation than congress. #tcot #semST	SEB	indirect	none	none
mr. t uses the #scientology method-never defend, always attack. #debatenight	CONS	indirect	against	none
I want a debate b/t Donald Trump and Hilary Clinton but it's their spouses on stage instead	MTSD	direct	none	favor

Table 6: **Examples of misclassifications by DSSD in different Trump datasets.** In SEB, most instances of 'none' do not mention Trump (example 1), while in the others like MTSD, these are tweets which mention him but do not express a clear favorable or unfavorable stance, which DSSD misclassifies as 'favor' or 'against' (example 3). On the other hand, it also misclassifies favorable or unfavorable tweets which indirectly mention Trump as 'none' (example 2).

Direct vs Indirect Stance. Recall that the advantage of stance detection over targeted sentiment detection is that in the former, *indirect* stance, where the target is not explicitly mentioned, can also be measured. Therefore, we compare the performance of DSSD for both direct and indirect stances in SEB and CONS in Figure 1.[14] We find that DSSD is better at measuring direct stances, especially those against the target, than indirect ones (c.f example 3 in Table 6) which corroborates previous findings of indirect stance being harder to automatically detect (Mohammad et al., 2017). Lower performance of automated methods for indirect stance, the advantage of stance detection over targeted sentiment analysis, implies a need for novel approaches.

No Stance. Figure 2 shows that DSSD misclassifies most and some portion of 'none' in non-SEB datasets, and SEB, respectively. This could be due to qualitative differences between the 'none' class in different datasets. From Table 3, we see that almost all tweets with no stance in SEB are of type indirect stance (example 1 in Table 6). [15] PRES and MTSD do not have instances of indirect stance and therefore tweets with no stance in them, directly mention Trump (example 2). DSSD misclassifying instances of no stance in PRES and MTSD, indicates that it does not recognize neutral mentions of targets as 'none'. The confusion be-

[14] MTSD and PRES, do not contain indirect stance.

[15] While example 1 seems unrelated to Trump, we argue that it still constitutes an indirect mention with no stance since all tweets in SEB contained stance-indicative or stance-neutral hashtags related to the target which were replaced with #semST during annotation by crowdworkets (Mohammad et al., 2016).

tween tweets which do not mention the target at all (tweets with indirect favorable or unfavorable stances) and tweets that mention the target but do not express a stance towards them (neutral tweets) could be due to the nature of weak labels used to train the method. **Our results indicate that the interplay of presence of stance, neutrality and directness needs to be investigated further.**

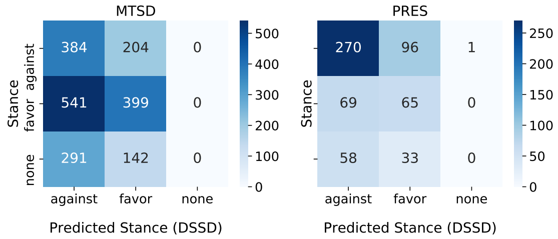

Figure 2: Confusion matrices of DSSD on MTSD and PRES. DSSD has high false negatives for 'none'.

6 Discussion

One of the goals of language technology, including NLP methods, is the ability to (re-)use them. Keeping in mind this vision, we investigate how an important construct in CSS, political approval, can be operationalized using existing NLP techniques, either through off-the-shelf sentiment and stance detection methods or through custom domain-specific methods. By comparing the performance of twelve methods over five datasets with approval towards seven targets, we find that targeted OTS methods do not perform well across targets or datasets that span over different time periods and have been collected using different collection strategies. Concretely, (i) targeted OTS methods do not generalize beyond the targets they were trained on. They are as good as or even worse than general-purpose lexicons in this case; (ii) even for familiar targets, targeted methods, especially stance detection, have high fluctuations and perform worse than sentiment lexicons for certain datasets; (iii) Finally, stance methods do not have a clear advantage over targeted sentiment in understanding approval due to the latter's low performance on indirect stance.

While researchers interested in measuring approval should use targeted constructs like stance or targeted sentiment instead of overall sentiment to avoid conceptual confusion, current targeted methods need to be improved before they can be used in an off-the-shelf manner. Since OTS targeted methods do not perform well for unknown targets, authors of papers on stance detection and target-dependent sentiment analysis should clarify if their method works only for certain targets (target-specific) or can be used to measure stance towards any unseen target (general-purpose), i.e., clarify the borders of their method's applicability. The high performance of sentiment lexicons, especially for unseen targets (Table 4), implies that these resources can be used with ML techniques for general-purpose stance detection. The poor performance of DSSD on other Trump datasets implies that, compared to sentiment analysis methods, stance methods are more susceptible to changes in topic and time. Future SemEval challenges should consider this when constructing test datasets and mention the hashtags and keywords they use for data collection. In our error analysis, we show that current stance detection methods, which are slated as being capable of measuring indirect opinions expressed via "pronouns, epithets, honorifics and relationships," perform poorly on indirect stance. This suggests that future research should explore approaches like coreference resolution (for pronouns), word sense disambiguation (for epithets), and background knowledge (relationships to other entities). Finally, to help practitioners and CSS researchers interested in measuring the approval of novel and familiar targets beyond a data collection setting familiar to an OTS method, we find that minimal in-domain models are preferable.

Limitations. This work does not capture all methods that have been proposed for assessing political approval but focuses on those that have been popular in the past or are exemplary for different types of methods (untargeted sentiment, targeted sentiment, and stance). Second, we only consider approval towards named entities, which we find is already a difficult task, especially for indirect stances. In the future, we hope to explore abstract topics like 'immigration' where differentiating between direct and indirect stance is non-trivial and ensemble models that combine the strengths of multiple methods.

Acknowledgement. We thank all the authors of datasets and methods used in this study for making their contribution available for reuse. We thank Mattia Samory, Katrin Weller, Juhi Kulshrestha and the anonymous reviewers for their constructive feedback and insightful discussions.

References

Nicholas Joseph Adams-Cohen. 2020. Policy change and public opinion: Measuring shifting political sentiment with social media data. *American Politics Research*, 48(5):612–621.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885, Austin, Texas. Association for Computational Linguistics.

Pablo Barberá. 2016. Less is more? how demographic sample weights can improve public opinion estimates based on twitter data. *Work Pap NYU*.

Esther van den Berg, Katharina Korfhage, Josef Ruppenhofer, Michael Wiegand, and Katja Markert. 2019. Not my president: How names and titles frame political figures. In *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*. Association for Computational Linguistics.

Manfred Max Bergman. 1998. A theoretical note on the differences between attitudes, opinions, and values. *Swiss Political Science Review*, 4(2):81–93.

Emily M Cody, Andrew J Reagan, Peter Sheridan Dodds, and Christopher M Danforth. 2016. Public opinion polling with twitter. *arXiv preprint arXiv:1608.02024*.

Raviv Cohen and Derek Ruths. 2013. Classifying political orientation on twitter: Its not easy! In *ICWSM*.

Frederick G Conrad, Johann A Gagnon-Bartsch, Robyn A Ferg, Michael F Schober, Josh Pasek, and Elizabeth Hou. 2019. Social media as an alternative to surveys of opinions about the economy. *Social Science Computer Review*, page 0894439319875692.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Sheridan Dodds, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher M Danforth. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PloS one*, 6(12):e26752.

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent Twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 49–54, Baltimore, Maryland. Association for Computational Linguistics.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Adam Hughes, Sono Shah, and Aaron Smith. 2020. Tweets by members of congress tell the story of an escalating covid-19 crisis. *Washington, DC: Pew Research Center*.

Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM*.

Kenneth Joseph, Lisa Friedland, William Hobbs, David Lazer, and Oren Tsur. 2017. ConStance: Modeling annotation contexts to improve stance classification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1115–1124, Copenhagen, Denmark. Association for Computational Linguistics.

Elena Kochkina, Maria Liakata, and Isabelle Augenstein. 2017. Turing at semeval-2017 task 8: Sequential approach to rumour stance classification with branch-lstm. *arXiv preprint arXiv:1704.07221*.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. Association for Computational Linguistics.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):1–25.

Micol Marchetti-Bowick and Nathanael Chambers. 2012. Learning for microblogs with distant supervision: Political forecasting with Twitter. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 603–612, Avignon, France. Association for Computational Linguistics.

Gregory E McAvoy. 2008. Substance versus style: Distinguishing presidential job performance from favorability. *Presidential Studies Quarterly*, 38(2):284–299.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41.

Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(3):26.

Joe Murphy, Craig A Hill, and Elizabeth Dean. 2014. Social media, sociality and survey research. *Social media, sociality, and survey research*, pages 1–33.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *ICWSM*.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2016. Social data: Biases, methodological pitfalls, and ethical boundaries. *Available at SSRN: https://ssrn.com/abstract=2886526 or http://dx.doi.org/10.2139/ssrn.2886526*.

Josh Pasek, Colleen A McClain, Frank Newport, and Stephanie Marken. 2019. Whos tweeting about the president? what big survey data can tell us about digital traces? *Social Science Computer Review*, page 0894439318822007.

Josh Pasek, H Yanna Yan, Frederick G Conrad, Frank Newport, and Stephanie Marken. 2018. The stability of economic correlations over time: Identifying conditions under which survey tracking polls and twitter sentiment yield similar conclusions. *Public Opinion Quarterly*, 82(3):470–492.

Jiaxin Pei, Aixin Sun, and Chenliang Li. 2019. Targeted sentiment analysis: A data-driven categorization. *arXiv preprint arXiv:1905.03423*.

Kevin M Quinn, Burt L Monroe, Michael Colaresi, Michael H Crespin, and Dragomir R Radev. 2010. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1):209–228.

Indira Sen, Fabian Floeck, Katrin Weller, Bernd Weiss, and Claudia Wagner. 2019. A total error framework for digital traces of humans. *arXiv preprint arXiv:1907.08228*.

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 551–557, Valencia, Spain. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Duyu Tang, Bing Qin, and Ting Liu. 2016. Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 214–224, Austin, Texas. Association for Computational Linguistics.

Mike Thelwall. 2017. The heart and soul of the web? sentiment strength detection in the social web with sentistrength. In *Cyberemotions*, pages 119–134. Springer.

Appendix for "On the Reliability and Validity of Detecting Approval of Political Actors in Tweets"

This appendix provides more details on the training data used for the custom methods (Appendix A), the evaluation datasets (Appendix B), and the training of different supervised methods including description of hyperparameters and how they were set (Appendix C).

A Training Data for Low-resource Custom Methods

We experiment with varying number of datapoints for training the low-resource custom methods and compare their performance against the OTS methods. The change in performance with increasing training data for Trump is in shown in Figure 3. We choose the least amount of data, 195 tweets, required to outperform the best OTS method, in this case, DSSD. With more training data, performance of customized methods improve but we attempt to show the least cost a researcher would incur for labeling additional data in their novel dataset for better performance than OTS methods. Custom methods for other targets also behave in a similar manner (c.f Figure 4), with certain targets like Putin outperforming the best OTS method, STB in this case, with fewer than 195 labeled tweets. Therefore, instead of having different training sizes for different targets, we use the same amount and find that the LR custom methods outperform OTS methods for all targets except Macron. The proportion of training data used for each target is mentioned in Table 9

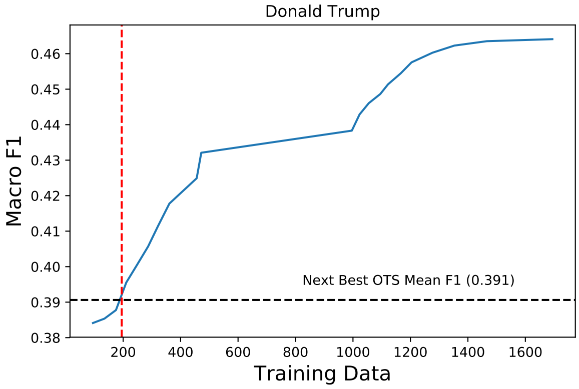

Figure 3: Relationship between increasing training data and performance (Mean Macro F1) for the target Donald Trump. We find that 195 datapoints are needed to train a custom model (LR in this case), that can outperform the best performing OTS method, DSSD.

B Evaluation Datasets

We briefly describe the datasets used for evaluation in Section 4.2. We provide more details on the specific datasets as well as how we rehydrated some of them (MTSD and PRES) based on tweet IDs released by the dataset authors (c.f Table 7). We also include the specific keywords and hashtags used to collect tweets and the period of data collection when available. The keywords used to collect the SEB data is not mentioned, and neither is the exact time period of data collection for SEB, SEA and MTSD therefore, based on the nature of the tweets, we estimate to be during the US 2016 elections.

C Training Supervised Methods

OTS Methods. The ML OTS methods we re-implement are SVM-SD, TD-LSTM, and DSSD. The training data used to re-implement these methods are described in Table 11. Since these methods are used in an off-the-shelf manner, we do not finetune them on a separate in-domain dev set. Nonetheless, the hyperparameters of TD-LSTM and DSSD are set according to the finetuning done on their original development set, while SVM-SD is finetuned through five-fold cross validation and grid search. The hyperparameters for these methods are listed in Table 10.

Dataset	Target	Stance			Methods Trained
		Against	Favor	None	
SemEval A (Train fold)	Atheism	304	92	117	SVM-SD DSSD
	Climate Change	15	212	168	
	Feminist Movement	328	210	126	
	Hillary Clinton	361	112	166	
	Legalization of Abortion	334	105	164	
Weak Labels Augenstein et al.	Donald Trump	5074	4645	8912	DSSD
Targeted Sentiment Dong et al.	miscellaneous	1411	1411	2826	TD-LSTM

Table 11: **Off-the-Shelf Methods' Training Datasets.** Training Datasets used for training various OTS supervised methods related to different named entity targets and their stance distribution. Note that the Hillary Clinton data is from the SEA training set and does not overlap with the test data in Table 3.

In-domain Methods. We train 4 types of in-domain models: Logistic Regression, Multinomial Naive Bayes, SVM and finetuned BERT. Since a researcher would train a model based on the novel target she wants to analyze, we train separate models for each target, leading to 28 different models (seven targets and 4 model types) over 5 runs. For

Dataset	**Keyword**	**Time Period**	**Original Size**	**Rehyd -rated Size**	**Data Decay (%)**	**Source**
SEB	not specified	∼2016 (pre-election)	707	707	0	http://saifmohammad.com/WebPages/StanceDataset.htm
SEA (HRC)	#GOHILLARY #WhyIAmNotVotingForHillary #hillary2016	∼2016 (pre-election)	294	294	0	
CONS	@realDonaldTrump, @HillaryClinton, Hillary, Clinton, Trump, Donald, #maga, #imwithher, #debatenight, #election2016, #electionnight	29th July to 7th November 2016	563	563	0	https://github.com/kennyjoseph/constance
MTSD	#DonaldTrump, #Trumpt,#Trump2016, #HillaryClinton, #Hillary, #Hillary2016	∼2016 (pre-election)	4455	3052	31.5	http://www.site.uottawa.ca/∼diana/resources/stance_data/
PRES	last-name, #first-name, first-name +(last-name/ country)	18th June to 30th August 2017	4200	3434	18.2	https://www.cl.uni-heidelberg.de/english/research/ downloads/resource_pages/TwitterTitlingCorpus/twitles.shtml

Table 7: **Specifications of Evaluation Datasets.** The datasets used for evaluating all off-the-shelf and custom methods, the keywords used to curate them, the period of data collection and source. We also include data decay rate of the two datasets we rehydrated due to some portion of tweets being deleted: MTSD and PRES.

Method	**Hyper -parameters**	**Hyper -parameters Bounds**	**Trump**		**Macron**		**Clinton**		**Zuma**		**Widodo**		**Erdoğan**		**Putin**	
			Values	**Train Time**	**Values**	**Train Time**	**Values**	**Train Time**	**Values**	**Train Time**	**Values**	**Train Time**	**Values**	**Train Time**	**Values**	**Train Time**
LR	C, penalty	[0.01, 0.1, 1, 10, 100] , [l1,l2]	10, l2	0.59	100, l2	0.51	1, l2	0.48	10, l2	0.44	1, l2	0.34	10, l2	0.44	10, l2	0.48
MNB	alpha	[0.001, 0.01]	0.01	0.11	0.01	0.11	0.01	0.12	0.01	0.11	0.001	0.08	0.001	0.1	0.001	0.1
SVM	C	[0.01, 0.1, 1, 10, 100]	0.1	0.25	10	0.16	0.1	0.22	0.1	0.21	0.01	0.18	0.1	0.17	0.1	0.17
BERT	Batch size, Learning rate (Adam), Number of epochs	N/A	32, 2e-5, 4	27.81	32, 2e-5, 4	20.8	32, 2e-5, 4	31.1	32, 2e-5, 4	22.4	32, 2e-5, 4	17.4	32, 2e-5, 4	22.2	32, 2e-5, 4	22.6

Table 8: Hyperparamters of the different custom methods used in this study.

Clinton	Erdoğan	Macron	Putin	Trump	Widodo	Zuma
13	32	35	31	5.5	46	31

Table 9: Proportion of training data used per target to train custom methods. We always use an absolute number of 195 tweets.

	hyperparameters	Values	Train Time
TD-LSTM	learning_rate, hidden layers, l2 regularization	0.01, 200, 0.001	609.9
SVM-SD	C	100	130.8
DSSD	learning_rate, batch size, epochs, hidden size	0.0001, 70, 4, 100	6167.2

Table 10: Hyperparamters of the ML OTS methods (SVM-SD, TD-LSTM and DSSD) used in this study.

each run, we use five-fold cross validation and grid-search to tune hyperparameters of LR, MNB and SVM, which are mentioned in Table 8. We use the default hyperparameters for finetuned BERT also included in the same table.

Compute Architecture. All models except BERT were trained or retrained on a 40 core Intel(R) Xeon(R) CPU E5-2690 (without GPU). All BERT models were finetuned on the custom data on Colab using a single Tesla P100-PCIE-16GB GPU. Run times (in seconds) for off-the-shelf and custom methods are included in Table 10 and 8, respectively.

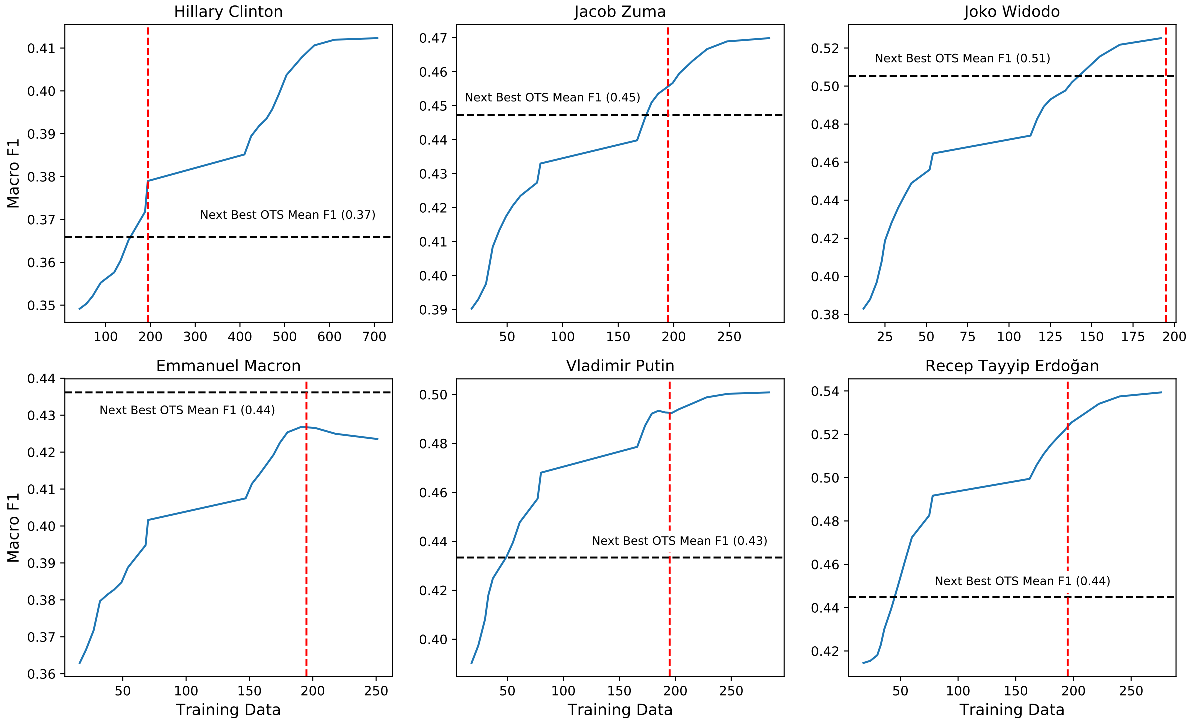

Figure 4: Relationship between increasing training data and performance (Mean Macro F1) for targets other than Trump.