A Few Topical Tweets are Enough for Effective User Stance Detection

Younes Samih
Dept. of Computational Linguistics,
University of Düsseldorf
Düsseldorf, Germany
Samih@phil.hhu.de

Kareem Darwish
Qatar Computing Research Institute, HBKU
Doha, Qatar
kdarwish@hbku.edu.qa

Abstract

User stance detection entails ascertaining the position of a user towards a target, such as an entity, topic, or claim. Recent work that employs unsupervised classification has shown that performing stance detection on vocal Twitter users, who have many tweets on a target, can be highly accurate (+98%). However, such methods perform poorly or fail completely for less vocal users, who may have authored only a few tweets about a target. In this paper, we tackle stance detection for such users using two approaches. In the first approach, we improve user-level stance detection by representing tweets using contextualized embeddings, which capture latent meanings of words in context. We show that this approach outperforms two strong baselines and achieves 89.6% accuracy and 91.3% macro F-measure on eight controversial topics. In the second approach, we expand the tweets of a given user using their Twitter timeline tweets, which may not be topically relevant, and then we perform unsupervised classification of the user, which entails clustering a user with other users in the training set. This approach achieves 95.6% accuracy and 93.1% macro F-measure.

1 Introduction

Stance detection entails identifying the position of a user towards a topic, an entity, or a claim (Mohammad et al., 2016b). Effective stance detection, particularly in the realm of social media, can be instrumental in gauging public opinion, identifying intersecting and diverging groups, and understanding issues of interest to different user communities (Magdy et al., 2016a). Much recent works have explored varying stance detection methods including supervised, semi-supervised, and unsupervised user classification (Darwish et al., 2020; Magdy et al., 2016a; Pennacchiotti and Popescu, 2011; Wong et al., 2013), and much of the work has fo-

Figure 1: Sample 2018 US midterm election related tweets that either express a very clear stance (a) or not (b)

cused on stance detection for Twitter users. The different approaches have advantages and disadvantages. For example, supervised methods are simple to implement, but they require manually annotated training data and their accuracy varies widely based on classification features, classification techniques, and the number of training and test examples (Magdy et al., 2016a). Though semi-supervised and unsupervised methods typically use user interactions and often may yield perfect classification, they are effective in classifying highly vocal users with many topical tweets (Darwish et al., 2020). Most of these methods produce sub-optimal results for users who rarely express their opinion, and for whom we may only have one or two topically related tweets. Though a single tweet might be explicitly clear, often it may lack sufficient context to determine the stance of the user. Figure 1 show two tweets that pertain to the 2018 US midterm elections, where the first expresses a lucid pro-Republican stance and the second could have been authored by a supporter of either the Republican or Democratic party. In this paper, we aim to effectively identify the stance of Twitter users towards specific targets (entities or topics), where the users have mentioned the targets in only a few tweets (less than two tweets on average).

To do so, we employ two approaches. In the first

approach, we classify users based on their tweets that are represented using contextualized embeddings, which capture latent meanings of words in context. Specifically, we use BERT embeddings to represent tweets, and we fine tune the embeddings for every topic. We compare this approach to two strong baselines, namely using Support Vector Machine (SVM) classification and fastText, which is a deep learning based classifier. In the second approach, we expand the tweets of a given user using their Twitter timeline tweets, and then we use the additional tweets, which would typically be not relevant to the topic at hand, to perform unsupervised classification of the user by clustering him/her with the users in the training set. Using such expansion allows us to make use of user homophily, which is manifested in the echo chambers that form on Twitter, where users with similar views tend to retweet similar accounts beyond the topic at hand. To test our approaches, we use a dataset containing tweets on 8 polarized US-centric topics. We also examine the effect of expansion when using SVM, fastText, and contextualized embeddings. For testing, we randomly selected 100 users for each topic that have less than 5 topical tweets, and we manually labeled them for stance. To construct the training set, we used unsupervised stance detection to automatically label the 5,000 most active users per topic, and for every topic we used a balanced set of 500 users per stance as our training set (Darwish et al., 2020). Since the approaches rely on different features and utilize different classification techniques, we indicate which approach works best under different conditions.

The contributions of this paper are as follows:

- We fine-tune contextualized embeddings to generate latent representations of tweets to effectively classify the stance of users based on only one or two tweets. We achieve an accuracy of 89.6% and macro F-measure of 91.3%, which are significantly higher than the scores achieved using two strong baselines.

- We show that using additional timeline tweets for the users that we wish to classify, and then using unsupervised classification, where we cluster the test user with users in our training, leads to an accuracy of 95.6% and macro F1-measure of 92.0%. In doing so, we extend prior work on unsupervised stance detection to effectively classify both users who are vocal on a topic as well as those with perhaps one or two topical tweets.

- We show that expanding user tweets using their timeline tweets can significantly improve some supervised classification setups.

- We conduct error analysis on our best setups to determine the sources of the errors and to guide the choice of classification methods.

- We plan to release the tweet IDs of the test set along with the associate gold labels. Further, we plan to release the code that performs classification based on contextualized embeddings.

2 Related Work

Over the last few years, much research has focused on user stance detection. The goal of stance detection is to ascertain the positions of users towards some target such as a topic, person, or claim (Thomas et al., 2006; Mohammad et al., 2016a; Barberá, 2015; Barberá and Rivero, 2014; Borge-Holthoefer et al., 2015; Cohen and Ruths, 2013; Colleoni et al., 2014; Conover et al., 2011; Fowler et al., 2011; Himelboim et al., 2013; Magdy et al., 2016a,b; Makazhanov et al., 2014; Weber et al., 2013). While stance may easily be detected by humans, machine learning models often fall short, particularly for users who talk about a target sparingly. Several studies have focused on modeling stance by introducing different features ranging from linguistic and structural features (Mohammad et al., 2016a) to network interactions and profile information (Borge-Holthoefer et al., 2015; Magdy et al., 2016a,b; Weber et al., 2013). Much work on stance detection involved using supervised and semi-supervised classification methods. One of the major downsides of both classification methods is the need for a seed list of manually labeled users, which is time consuming and requires topic expertise. Supervised learning is sensitive to the classification features, the size of the training sets, the number of available tweets for users in the test set, and the classification algorithm (Borge-Holthoefer et al., 2015). Some common classification features include: lexical, syntactic, and semantics feature; network features such as retweeted accounts and user mentions; content features such as words and hashtags; and user profile information such as name and location (Aldayel and Magdy, 2019; Magdy et al., 2016a,b; Pennacchiotti and Popescu, 2011). Some commonly used classification algorithms include SVMs and deep learning classification (Zarrella and Marsh, 2016). Popat et al.

(2019) presented a neural network model for stance classification by augmenting BERT representations with a novel consistency constraint to determine stance with respect to both a claim and perspective. We extend their work in two ways, namely: we drop the need to have a claim and perspective, and we couple BERT supervised classification with unsupervised classification to effectively tag vocal and non-vocal users. Semi-supervised methods such as *label propagation* (Barberá, 2015; Borge-Holthoefer et al., 2015; Weber et al., 2013) often rely on two users retweeting identical accounts or tweets to propagate a label of one user to another. Though such typically achieves high precision (often above 95%) (Darwish et al., 2018), it is generally successful in tagging vocal users with strong opinions. Recently, Darwish et al. (2020) have introduced a highly effective unsupervised method for predicting the stance of prolific Twitter users towards controversial topics. By projecting users onto a low-dimensional space and then clustering them allows for clear separation between vocal users with respect to their stance (Darwish et al., 2020). This method confers two main advantages over previous methods, namely: it does not require any initial manual labeling, and classification accuracy is nearly perfect. However, it is successful in labeling vocal users only and fails on users with very few topical tweets. We extend prior work on unsupervised stance detection to effectively classify both prolific and non-prolific users in a holistic way by aggregating both supervised and unsupervised methods. Further, we extend prior deep-learning based supervised classification to use contextual embeddings that capture syntactic and semantic features of words in context.

We are framing the problem as user-based classification, which is common in the computational social science community, as opposed to tweet-based classification, which is common in the NLP community (Mohammad et al., 2016b). This is motivated by two aspects, namely: 1) tweets often don't provide sufficient context for proper annotation; and 2) users have durable stances over time. For example, if someone says "Most important election in history! Vote!", it is nearly impossible to know if the author's position without context.

3 Data Sets

Topics   Our dataset includes tweets on eight polarizing topics that are US-centric, which were gra-

ciously provided to us by Stefanov et al. (2020). Table 1 lists all the topics including when the tweets were collected and the number of tweets per topic. The topics include both long-standing issues such as gun control and transient issues such as the nomination of Judge Kavanaugh to the US Supreme Court. There is also a non-political issue, namely vaccination. The tweets were also filtered based on user-stated locations to limit the data to US users. The filtering was done using a gazetteer that includes either US (or its variants) and state names (and their abbreviations).

Topic	Date Range	Tweets
Climate change	Feb 25–Mar 4, 2019	1,284,902
Gun control	Feb 25–Mar 3, 2019	1,782,384
Ilhan Omar (remarks on Israeli lobby)	Mar 1–9, 2019	2,556,871
Immigration	Feb 25–Mar 4, 2019	2,341,316
Midterm (elections 2018)	Feb 25–Mar 3, 2019	2,564,784
Kavanaugh (nomination to Supreme Court)	Sept. 28-30 & Oct. 6-9, 2018	2,322,141
Vaccination	Mar 1–9, 2019	301,209

Table 1: Controversial topics used in study.

Training Set   Given the tweets for every topic, we performed per topic unsupervised stance detection (Darwish et al., 2020). This approach identifies the most active n users per topic and computes similarity between them based on a common feature, such as which hashtags they use or which accounts that they retweet. Next, the users are projected onto a lower dimensional space in a manner where similar users are brought closer together and dissimilar users are pushed further apart. Then the projected users are clustered. Using the best reported setup of Darwish et al. (2020), we used the 5,000 most active users with at least 10 tweets, computed similarity between them based on which accounts they retweeted, projected users using UMAP (McInnes and Healy, 2018), and clustered them using the mean shift clustering algorithm (Fukunaga and Hostetler, 1975). Stefanov et al. (2019) estimated the accuracy of the unsupervised approach on the 8 topics to be 98%. Next, we took 500 random users from the two largest clusters to construct a balanced training set, and we manually inspected a few users from each cluster to give an overall label to each cluster (ex. pro- or anti- gun control). Further, we crawled the timeline tweets of the users in our training set.

Topic	Skip	Agree	Pro	Anti
Climate	10.6	100	81	18
Gun control	71.0	85	62	37
Ilhan	1.0	95	52	47
Immigration	10.7	90	57	42
Midterm	0.0	90	71	28
Police/Racism	0.0	100	80	20
Kavanaugh	2.0	100	39	60
Vaccine	1.0	100	90	9
Avg.	12.0	95	66.5	32.6

Table 2: Test set information: % skipped tweets, inter-annotator agreement, % pro, and % anti.
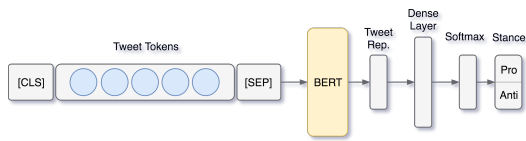

Figure 2: Fine-tuning BERT for stance classification.

Test Set For each topic, we randomly selected 200 users who have less than 5 tweets. The average number of tweets per user ranged across topics between 1.25 and 1.77 tweets. An annotator who is well versed with US politics manually examined the per topic tweets of users to determine their stances. If the tweets of a user were not sufficient to ascertain their stance, the annotator manually searched and examined their tweets on Twitter in an effort to find further clues. If no conclusive evidence of stance were found, the annotator skipped the user. The annotator labeled up to 100 users per topic. We asked another annotator to annotate 20 tweets per topic to ascertain inter-annotator agreement. Table 2 lists the percentage of skipped tweets, inter-annotator agreement, and percentage of pro- and anti-tweets. Next, we scraped the timeline of all the labeled users. Due to the time difference between collecting topical tweets and when we initiated the scraping of users' timelines, some user accounts were deleted, suspended, or made protected. Table 3 lists the number of labeled users and the subset of them for whom we were able to scrape their timelines. Thus, we put users for whom we were not able to collect timeline tweets into Set A, and we put the remaining users in Set B. We report results for both sets separately.

4 Classification Models

Supervised Classification As baselines, we used two different classification methods, namely

Topic	Labeled Users	w/Timelines
Climate change	100	54
Gun control	58	26
Ilhan Omar	100	39
Immigration	100	43
Midterm	100	45
Police/Racism	100	54
Kavanaugh	100	55
Vaccines	100	57

Table 3: Per topic labeled users in test set along with the number of users for which were able to scrape their timelines

a Support Vector Machines (SVM) classifier and a deep learning based text classifier. For the SVM classifier, we used the SVM^{Light} implementation with a linear kernel with default parameters (Joachims, 2002)[1]. We employed two feature types, namely: the accounts that users retweeted; and the words in tweets, including retweeted accounts, hashtags, and user mentions and replies. Prior work has shown that using retweeted accounts as features yields better results compared to using the content of tweet (Darwish et al., 2018). When using words in tweets, we tokenized tweets using NLTK (Bird et al., 2009), removed all URLs and emoticons, retained all hashtags and user mentions, and specifically delineated retweeted accounts by adding 'RT_' before them. We chose to distinguish between retweeted accounts and user mentions because retweeting commonly signifies agreement and user mentions (including replies) may indicate opposition. We concatenated the aforementioned features from all the tweets of a user, and we constructed a feature vector, where the value of each unique feature was set to its frequency across all tweets of a user. For the deep learning based classifier, we used fastText, which is an efficient text classifier that has been shown to be effective for different text classification tasks (Joulin et al., 2016). Since fastText was designed for sentence-level classification, we opted to perform tweet-level classification. During training, we assigned the label of a user to all his/her tweets. During testing, we averaged per class confidence scores across all tweets for a user, and we assigned the label with the highest average confidence to the user. As for features, we used all the words in tweets, and we preprocessed tweets in the manner described earlier for

---

[1] http://svmlight.joachims.org/

SVM. We opted not use retweeted accounts only as the number of retweeted accounts was arbitrary for each user and fastText is not well suited for long input text.

Contextualized Embeddings   Over the last several years, pre-trained embedding (Mikolov et al., 2013; Pennington et al., 2014) have helped achieve significant improvements in a wide range of classification tasks in natural language processing. Representing words as vectors in a low-dimensional continuous space and then using them for downstream tasks lowered the need for extensive manual feature engineering. However, these pre-trained vectors are static and fail to handle polysemous words, where different instances of a word have to share the same representation regardless of context. More recently, different deep neural language models have been introduced to create contextualized word representations that can cope with the issue of polysemy and the context-dependent nature of words. Models such as OpenAi GPT (Radford et al., 2018), ELMo (Peters et al., 2018), BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019), and UMLFIT (Howard and Ruder, 2018), to name a few, have achieved groundbreaking results in many NLP classification and language understanding tasks. For this paper, we use $BERT_{base-multilingual}$[2] (referred to hereafter simply as BERT), which we fine-tune for stance detection, as this eliminates the need for heavily engineered task-specific architectures. BERT is pre-trained on Wikipedia text from 104 languages and comes with hundreds of millions of parameters. It contains an encoder with 12 Transformer blocks, hidden size of 768, and 12 self-attention heads. As shown in Fig. 2, We fine-tuned BERT by adding a fully-connected dense layer followed by a softmax output layer, minimizing the binary cross-entropy loss function for the training data. For all experiments, we used HuggingFace[3] transformer implementation with PyTorch[4] as it provides pre-trained weights and vocabulary. As for features, we used all the words in tweets that were preprocessed in the manner described earlier for SVM and fastText. Similar to fastText, we performed tweet-level clas-

sification, and we used the average softmax output scores per class across all tweets for a user to assign a label to a test user.

Unsupervised Classification   For unsupervised classification, we used the same unsupervised classification method described earlier, which we used to prepare the training set. Specifically, we constructed a feature vector for each test user based on the accounts he/she retweeted, computed its similarity to all users in the training set, projected all the users in the training along with the test user into a lower dimensional space using UMAP, and lastly clustered the users using mean shift. We then labeled the test user using the majority label of the cluster in which the user appeared.

5   Experiments

We split users in our test set on the basis of whether we were able to crawl their timelines or not. Set A includes users for which we were not able to obtain their timeline tweets. Set B includes users for which we were able to collect their timeline tweets. We separated between them, because Set B would allow us to compare between setups that use timeline tweets with those that do not on identical users.

For Set A, we always trained on the training users with their on-topic tweets and the tested on the test users, who typically had less than 2 tweets on average. We used four different classification setups, namely using fastText, SVM with retweeted accounts as features (SVM_{RT}), SVM with all words as features (SVM_{TEXT}), and fine-tuned BERT embeddings with a dense neural layer and softmax output (BERT). We experimented with using the unsupervised method on Set A, but the unsupervised algorithm was not able to assign any test user to a cluster, mostly because the number of tweets and subsequently retweeted users per test user were too few. For Set B, we experimented with the same classifiers using four different conditions, namely: not expanding either training or test sets with users' timeline tweets; expanding the test set only; expanding the training set only; and expanding both the training and test sets.

6   Results and Discussion

For all experiments, we report on per topic accuracy ($\mathcal{A}$) and macro precision ($\mathcal{P}$), recall ($\mathcal{R}$), and F-measure ($\mathcal{F}$) across stances on a topic. Table 4

---

[2]We also experimented with different contextualized embedding, such as RoBERTa (Liu et al., 2019), Albert (Lan et al., 2019), and XLM (Lample and Conneau, 2019) and $BERT_{base-multilingual}$ performed the best.

[3]https://github.com/huggingface/transformers

[4]https://pytorch.org/

	fastText				SVM$_{RT}$				SVM$_{TEXT}$				BERT			
Topic	$\mathcal{A}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	$\mathcal{A}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	$\mathcal{A}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	$\mathcal{A}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$
Kavanaugh	83.5	83.2	83.2	83.2	73.0	**83.3**	70.6	69.2	76.0	76.3	71.5	72.4	**84.7**	81.6	**83.8**	**82.7**
Vaccine	88.7	73.1	93.8	78.2	**88.2**	44.1	50.0	46.9	87.0	43.5	50.0	46.5	85.7	**98.0**	**85.7**	**91.4**
Ilhan	87.0	87.7	86.4	86.7	65.1	79.7	64.3	59.5	52.4	26.2	50.0	34.4	**87.9**	**91.1**	**86.4**	**88.7**
Gun Control	92.3	91.8	90.9	91.3	65.4	75.0	73.5	65.3	72.7	79.6	77.5	72.6	**93.8**	**97.6**	**93.0**	**95.2**
Police Racism	94.7	89.8	94.0	91.7	94.9	97.1	85.7	90.2	83.0	41.5	50.0	45.4	**96.0**	**98.3**	**96.7**	**97.5**
Climate Change	95.8	93.8	95.3	94.5	82.9	90.9	62.5	65.0	81.6	90.2	62.5	64.6	**95.7**	**96.3**	**98.1**	**97.2**
Midterm	85.9	83.5	84.1	83.8	87.3	87.4	82.7	84.4	69.8	75.6	77.9	69.7	**90.2**	**92.2**	**93.7**	**92.9**
Immigration	84.4	83.8	84.0	83.	65.3	81.9	55.3	48.5	59.0	29.5	50.0	37.1	**89.5**	**91.2**	**91.2**	**91.2**
Average	89.0	85.8	88.9	86.6	77.8	79.9	68.1	66.1	72.7	57.8	61.2	55.3	**90.4**	**93.3**	**91.1**	**92.1**

Table 4: Results on Sets A (no expansion). The best results in a row are in bold.

reports the results on Sets A where we were not able to expand the test set using timeline tweets. As the results show, BERT yielded the best results in terms of $\mathcal{A}$, $\mathcal{P}$, $\mathcal{R}$, and $\mathcal{F}$ for most topics, with the highest overall averages across all scores. fastText trailed BERT, and SVM$_{TEXT}$ performed much worse. SVM$_{RT}$ performed better than SVM$_{TEXT}$. This suggests that BERT, which uses contextual embeddings, is effective in performing accurate stance detection, even when classifying users with a very small number of topical tweets. As for the Unsupervised method, using the unsupervised method was not able to assign any test user to a cluster, mostly because the number of tweets per test user were too few. Hence, we omitted the unsupervised method from Table 4. Table 5 shows the results on Set B, where we expanded the test, training, or either or both training and test user tweets using timeline tweets. The results suggest the following:

- For BERT and fastText, which rely on the content of the tweets, we achieved the best results with no expansion or when we only expanded the training set. The inclusion of non-topical tweets in the test set led to worse results overall. We suspect that is happened because of the mismatch between the training and test sets.

- For SVM$_{RT}$ and Unsupervised classification, which rely exclusively on whom users retweeted, the expansion of the test dramatically improved overall $\mathcal{A}$, $\mathcal{P}$, $\mathcal{R}$, and $\mathcal{F}$. The positive improvement for both after timeline expansion suggests that the accounts that a user retweets are a strong signal of stance across multiple topics, and stances on multiple topics are likely correlated. For example, a user who supported the Kavanaugh nomination was likely to vote republican in the midterm elections. For future work, we plan to examine cross topic classification.

- Similar to the results observed for Set A (4), when no expansion is used, BERT led to the best overall results. However, using unsupervised classification led to the best overall results across all setups, with expanding the test set only yielding slightly better results than expanding both the training and test sets. Expanding the test set only is significantly more efficient than expanding both training and test sets.

- Using unsupervised classification failed to all users in the test set for any topic when the test set was not expanded, mostly because the number of tweets per test user and subsequent number of retweeted accounts were too few.

- SVM$_{TEXT}$ yielded the worse results overall, despite the inclusion of all the features in the tweets, such as retweeted accounts, hashtags, words, etc. It seems that the inclusion of more features (compared to SVM$_{RT}$) confused the classifier leading to lower results.

- SVM$_{TEXT}$ and SVM$_{RT}$ led to the lowest results when we only expanded the training set. For both setups, the classifier classified all users as belonging to one of the stances or the other. Hence, $\mathcal{R}$ for one class was 100.0 and 0.0 for the other (with macro $\mathcal{R} = 50.0$). We suspect that expanding the feature space in the training set confused the SVM classifier. Both setups are unusable.

We computed the standard deviation (SD) of all our measures across topics for every setup. Lower SD coupled with high $\mathcal{A}$ and $\mathcal{F}$ is desirable as they indicate the setup produces consistently high results across topics. Unsupervised classification yielded the lowest SD values and highest overall score. BERT and fastText with no expansion and SVM$_{RT}$ with expanded test set had slightly higher SD. Thus, if we are able to scrape a user's timeline tweets, it is advantageous to use a method that relies on which accounts a user retweets, with unsupervised classification producing the best results. As we will show in the error analysis, the success of

fastText

Topic	No Expansion $\mathcal{A}$	P	R	F	Expanded Test only $\mathcal{A}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	Expanded Test and Train $\mathcal{A}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	Expanded Train only $\mathcal{A}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$
Kavanaugh	**82.8**	**82.5**	**82.3**	**82.4**	57.9	58.8	59.1	57.7	70.2	69.5	70.1	69.6	80.2	79.9	80.1	80.0
Vaccine	**86.4**	**68.3**	**85.0**	**72.3**	25.6	51.0	54.1	23.0	69.6	53.8	70.3	49.1	82.0	52.2	52.7	52.4
Ilhan	**85.9**	**89.4**	**85.3**	**85.4**	54.8	48.9	50.0	35.5	73.8	76.2	72.0	71.9	84.4	84.3	84.4	84.3
Gun Control	70.3	72.1	76.2	69.5	46.6	49.0	48.9	46.3	69.2	68.6	70.4	68.3	**77.5**	**77.5**	**78.6**	**76.0**
Police Racism	**88.0**	81.8	90.3	84.5	50.9	49.5	49.2	45.5	83.0	74.7	78.9	76.3	**88.0**	**81.0**	**83.1**	**82.0**
Climate Change	75.6	68.1	79.6	68.0	78.2	51.3	51.0	50.9	80.3	68.6	80.3	71.0	**87.6**	**80.7**	**87.6**	**83.2**
Midterm	84.9	79.4	82.2	80.6	42.6	49.5	49.3	41.7	81.0	74.5	77.4	75.7	**91.7**	**91.0**	**88.5**	**89.6**
Immigration	87.5	87.7	87.7	87.5	49.7	50.1	50.1	49.5	79.0	78.4	79.0	78.6	83.0	84.3	84.3	83.0
Average	82.6	**78.6**	**83.5**	**78.9**	50.8	51.0	51.5	43.8	65.8	70.5	74.8	70.1	**84.3**	**78.6**	79.9	**78.8**
Std Dev	6.0	7.8	4.2	7.0	13.9	3.1	3.3	9.9	5.3	7.2	4.2	8.6	4.3	10.8	10.8	10.6

SVM$_{RT}$

Topic	No Expansion $\mathcal{A}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	Expanded Test only $\mathcal{A}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	Expanded Test and Train $\mathcal{A}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	Expanded Train only $\mathcal{A}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$
Kavanaugh	73.7	78.9	72.5	71.7	84.9	84.9	83.8	84.2	**86.8**	**88.7**	**84.8**	**85.8**	52.6	26.3	50.0	34.5
Vaccine	95.5	47.7	50.0	48.8	**96.4**	**98.2**	**75.0**	**82.4**	**96.4**	**98.2**	**75.0**	**82.4**	95.5	47.7	50.0	48.8
Ilhan	57.1	77.3	55.9	45.8	**89.5**	**91.7**	**88.9**	**89.2**	71.1	82.3	69.4	67.2	51.4	25.7	50.0	34.0
Gun Control	66.7	72.7	76.9	66.3	75.0	**78.6**	**81.3**	74.8	**79.2**	76.9	75.0	**75.8**	72.2	36.1	50.0	41.9
Police Racism	90.6	94.6	78.6	83.5	92.3	88.0	**92.1**	**89.7**	90.4	**94.4**	79.2	83.9	78.1	39.1	50.0	43.9
Climate Change	92.3	96.0	70.0	76.5	_100.0_	_100.0_	_100.0_	_100.0_	96.2	97.9	85.7	90.6	87.2	43.6	50.0	46.6
Midterm	**95.5**	**97.2**	90.0	**93.0**	92.9	89.3	91.9	90.5	88.1	84.3	81.9	83.0	77.3	38.6	50.0	43.6
Immigration	59.5	78.6	55.9	46.9	**97.6**	**97.5**	_97.8_	**97.6**	73.8	79.4	71.5	71.0	54.1	27.0	50.0	35.1
Average	78.9	80.4	68.7	66.6	**91.1**	**91.0**	**88.8**	**88.6**	85.2	87.8	77.8	80.0	71.0	35.5	50.0	41.0
Std Dev	15.4	15.3	12.8	16.7	7.5	6.9	7.9	7.6	9.0	7.8	5.7	7.4	15.7	7.8	0.0	5.4

SVM$_{TEXT}$

Topic	No Expansion $\mathcal{A}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	Expanded Test only $\mathcal{A}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	Expanded Test and Train $\mathcal{A}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	Expanded Train only $\mathcal{A}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$
Kavanaugh	68.5	67.8	67.0	67.2	**70.9**	**71.3**	**69.0**	**69.2**	56.4	65.9	60.4	53.8	42.6	21.3	50.0	29.9
Vaccine	**92.9**	**46.4**	50.0	**48.2**	89.5	46.4	48.1	47.2	49.1	52.9	**61.1**	40.2	7.1	3.6	50.0	6.7
Ilhan	52.6	26.3	50.0	34.5	53.9	26.9	50.0	35.0	71.8	72.2	72.2	71.8	47.4	23.7	50.0	32.1
Gun Control	56.0	72.5	65.6	54.8	69.2	**76.5**	**76.5**	69.2	46.2	59.1	56.2	44.9	36.0	18.0	50.0	26.5
Police Racism	77.4	38.7	50.0	43.6	**85.2**	**81.0**	72.6	**75.5**	66.7	61.9	66.7	61.4	22.6	11.3	50.0	18.5
Climate Change	92.5	96.0	71.4	77.9	79.6	43.0	45.7	44.3	42.6	59.2	67.0	41.0	13.2	6.6	50.0	11.7
Midterm	61.4	68.5	75.0	60.4	80.0	71.3	62.1	64.0	**84.4**	**78.3**	**86.4**	**80.6**	22.7	11.4	50.0	18.5
Immigration	54.8	27.4	50.0	35.4	58.1	62.1	53.2	45.0	**86.1**	**87.7**	**84.8**	**85.4**	54.8	27.4	50.0	35.4
Average	69.5	55.5	59.9	52.7	**73.3**	**59.8**	**59.7**	**56.2**	62.9	67.2	69.3	59.9	30.8	15.4	50.0	22.4
Std Dev	15.4	23.1	10.2	14.3	11.8	17.8	11.2	14.0	15.9	10.8	10.4	16.7	15.9	7.9	0.0	9.5

BERT

Topic	No Expansion $\mathcal{A}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	Expanded Test only $\mathcal{A}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	Expanded Test and Train $\mathcal{A}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	Expanded Train only $\mathcal{A}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$
Kavanaugh	**82.5**	**79.**	**81.8**	**80.7**	60.9	51.1	69.5	58.9	70.9	62.2	71.0	66.3	80.5	76.9	78.9	77.9
Vaccine	89.1	**99.0**	88.7	93.6	38.9	96.7	37.5	54.0	73.5	98.3	73.6	84.2	90.9	98.2	**91.7**	**94.8**
Ilhan	91.6	**93.6**	90.7	92.1	61.7	62.7	74.4	68.1	76.0	72.9	89.5	80.4	93.0	92.1	**94.6**	**93.3**
Gun Control	**86.2**	88.9	**95.0**	**83.5**	88.8	70.8	60.1	65.0	70.4	83.9	67.3	74.7	73.0	90.0	69.2	78.3
Police Racism	**93.0**	96.7	94.4	**95.5**	67.6	86.0	70.7	77.6	83.4	93.6	84.9	89.0	91.6	93.9	**95.4**	94.7
Climate Change	**94.1**	**97.4**	**95.0**	**96.2**	89.0	75.0	75.0	81.4	71.0	92.8	71.6	80.8	80.5	93.3	82.4	87.5
Midterm	91.7	93.3	95.1	94.2	57.8	81.6	57.9	67.7	82.3	92.2	84.0	87.9	**94.3**	**95.1**	**97.5**	**96.3**
Immigration	88.1	92.4	86.0	89.1	65.3	72.9	65.3	68.9	81.3	84.3	83.8	84.0	**92.6**	**97.4**	**88.4**	**92.7**
Average	**89.6**	**93.4**	**89.4**	**91.3**	62.4	74.6	63.8	67.7	76.1	85.0	78.2	80.9	87.0	92.1	87.3	89.4
Std Dev	8.8	7.2	11.4	8.6	12.9	13.1	11.5	8.4	5.1	11.4	7.7	7.0	7.4	6.3	9.1	7.0

Unsupervised

Topic	No Expansion				Expanded Test only $\mathcal{A}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	Expanded Test and Train $\mathcal{A}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	Expanded Train only			
Kavanaugh					84.6	84.2	84.2	84.2	**90.4**	**89.2**	**91.3**	**89.9**				
Vaccine					**96.3**	83.3	**99.0**	**89.5**	95.3	97.1	75.0	81.8				
Ilhan					**91.9**	91.6	92.1	91.8	**91.9**	91.6	92.1	91.8				
Gun Control					_95.8_	_90.6_	_86.4_	_86.9_	87.5	90.6	86.4	86.9				
Police Racism					_100.0_	_100.0_	_100.0_	_100.0_	98.0	98.7	95.8	97.2				
Climate Change					96.2	97.8	88.9	92.6	92.3	96.7	85.0	89.5				
Midterm					_100.0_	_100.0_	_100.0_	_100.0_	95.0	96.8	90.9	93.3				
Immigration					_100.0_	_100.0_	_100.0_	_100.0_	97.6	97.8	97.5	97.6				
Average					_95.6_	93.4	_93.8_	_93.1_	93.5	_94.8_	89.3	91.0				
Std Dev					4.9	6.6	6.3	5.9	3.4	3.5	6.7	4.9				

Table 5: Results on Set B with and without expansion of either training or test sets. Highest $\mathcal{A}$, $\mathcal{P}$, $\mathcal{R}$, and $\mathcal{F}$ per method are bolded, and highest values overall are underlined. The table reports the average and standard deviation.

Error Type	No.	Examples
Unexplained	52	*Ilhan*: RIP CONS ... When they see me they take their hat off... I am DJ
		Gun Control: @BreitbartNews @NRA more lies
		Climate Change: $200 million a year could reverse climate change
Vague	48	*Kavanaugh*: following the senate confirmation vote
		Midterm: RT_@ali irresponsible for twitter
		Police & Racism: after 40 yrs of reflection
Quoting other side	24	*Immigration*: this is a real crisis at the border
		Kavanaugh: why did jeff flake demand an investigation and then accept a bogus one
Sarcasm	4	*Immigration*: RT_@infantry0300 someone should let "Ms Hitler was a really great guy until he crossed the border into Poland"

Table 6: BERT error types with examples

unsupervised classification is contingent on users retweeting a sufficient number of times, particularly politically related accounts in our case. When timeline tweets are not available, it is best to use contextualized embeddings.

Error Analysis We analyzed all the errors in Set B that were produced by BERT with no expansion, as it represents the best results when expansion is not possible, and those produced by unsupervised classification with the expansion of the test set only, as this produced the best overall results. Since we used BERT to perform tweet-level classification, we manually inspected all 129 misclassified tweets across all topics. Generally we found four types of errors, namely: unexplainable errors where the tweets clearly expressed stance, but the classifier mislabeled them; vague tweets that have no clear clues; tweets in which the user uses the language of the opposing side; and sarcastic tweets. Table 6 lists the error types with their frequencies and provides example tweets. When we used LIME (Ribeiro et al., 2016) to analyze the output of BERT, we noticed two important phenomena. First, BERT was able to identify stance based on retweeted accounts, and not just the text of the tweet. Second, BERT was able to learn correlations between topic. For example, the tweet "RT **BernieSanders**: I believe **health care** must be a right", BERT based its decision on BernieSanders, the democratic presidential hopeful in 2016 and 2020, and on "health care", where positions on *health care* and *climate change* are often aligned.

For unsupervised classification, we manually examined all 15 users that were missclassified, of whom 7 were from the Kavanaugh topic. Prominent reasons for incorrect classification were: Lack of sufficient retweets for a user, where the percentage of retweets ranged between 1-1.4% of all tweets for three of the misclassified users (2 for *climate change* and 1 for *Kavanaugh*); Geo-

graphic mislabeling, where 2 accounts were not US accounts (1 for *Ilhan* and 1 for *Kavanaugh*); Users retweeting mostly apolitical accounts such as music, art, or cars related accounts (1 for *climate change*, 1 for *vaccine*, and 3 for *Kavanaugh*) – retweeting of politically biased accounts and media sources seem to provide strong signals for classification; or User goes against the general opinion of his group as in the clearly republican user who was criticizing the National Rifle Association (NRA) (*gun control*).

Thus, the most common reason for misclassification was the dearth of retweets from politically oriented or topically related accounts.

7 Conclusion

In this paper, we presented two methods for classifying users according to their stance towards a target. The first utilizes contextualized embeddings to represent tweets and then uses deep neural network for classification. This approach led to results that outperform two strong baselines. The second utilizes additional tweets from users' timelines to cluster test users with other users with known labels in an unsupervised manner. The first method yielded the best results when timeline tweets were not available, while the second yielded even better results overall. Given the overall setup described in the paper, where the training data was obtained using unsupervised user classification, we can automatically label the most active users with nearly perfect accuracy, and we can label users with only few topical tweets with high accuracy, often above 95% when we can obtain their timeline tweets. For future work, we plan to explore the effectiveness of cross topic classification, where training and testing are done on different topics. Perhaps, we can build unified models that could be used across multiple topics for a given a population of Twitter users (ex. users who are interested in US politics).

References

Abeer Aldayel and Walid Magdy. 2019. Your stance is exposed! analysing possible factors for stance detection on social media. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Pablo Barberá. 2015. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis*, 23(1):76–91.

Pablo Barberá and Gonzalo Rivero. 2014. Understanding the political representativeness of Twitter users. *Social Science Computer Review*, page 0894439314558836.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Javier Borge-Holthoefer, Walid Magdy, Kareem Darwish, and Ingmar Weber. 2015. Content and network dynamics behind Egyptian political polarization on Twitter. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 700–711.

Raviv Cohen and Derek Ruths. 2013. Classifying political orientation on Twitter: It's not easy! In *ICWSM 2013*.

Elanor Colleoni, Alessandro Rozza, and Adam Arvidsson. 2014. Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of Communication*, 64(2):317–332.

Michael D Conover, Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. 2011. Predicting the political alignment of Twitter users. In *Proceedings of the 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom)*, pages 192–199.

Kareem Darwish, Walid Magdy, Afshin Rahimi, Timothy Baldwin, Norah Abokhodair, et al. 2018. Predicting online islamophobic behavior after# parisattacks. *The Journal of Web Science*, 4(3):34–52.

Kareem Darwish, Peter Stefanov, Michaël Aupetit, and Preslav Nakov. 2020. Unsupervised user stance detection on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 141–152.

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.

James H Fowler, Michael T Heaney, David W Nickerson, John F Padgett, and Betsy Sinclair. 2011. Causality in political networks. *American Politics Research*, 39(2):437–480.

Keinosuke Fukunaga and Larry Hostetler. 1975. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on information theory*, 21(1):32–40.

Itai Himelboim, Stephen McCreery, and Marc Smith. 2013. Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on Twitter. *Journal of Computer-Mediated Communication*, 18(2):40–60.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia. Association for Computational Linguistics.

T. Joachims. 2002. *Learning to Classify Text Using Support Vector Machines – Methods, Theory, and Algorithms*. Kluwer/Springer.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Walid Magdy, Kareem Darwish, Norah Abokhodair, Afshin Rahimi, and Timothy Baldwin. 2016a. #isisisnotislam or #deportallmuslims?: Predicting unspoken views. In *Proceedings of the 8th ACM Conference on Web Science*, pages 95–106. ACM.

Walid Magdy, Kareem Darwish, and Ingmar Weber. 2016b. #failedrevolutions: Using Twitter to study the antecedents of ISIS support. *First Monday*, 21(2).

Aibek Makazhanov, Davood Rafiei, and Muhammad Waqar. 2014. Predicting political preference of Twitter users. *Social Network Analysis and Mining*, 4(1):1–15.

Leland McInnes and John Healy. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016a. 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, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016b. Semeval-2016 task 6: Detecting stance in tweets. In *SemEval 2016*, volume 16.

Marco Pennacchiotti and Ana-Maria Popescu. 2011. Democrats, Republicans and Starbucks afficionados: user classification in Twitter. In *Proceedings of ACM SIGKDD*, pages 430–438.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *ArXiv*, abs/1802.05365.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2019. STANCY: Stance classification based on consistency cues. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6412–6417, Hong Kong, China. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144.

Peter Stefanov, Kareem Darwish, Atanas Atanasov, and Preslav Nakov. 2019. Predicting the topical stance of media and popular twitter users. *arXiv preprint arXiv:1907.01260*.

Peter Stefanov, Kareem Darwish, Atanas Atanasov, and Preslav Nakov. 2020. Predicting the topical stance and political leaning of media using tweets. *The 58th Annual Meeting of the Association for Computational Linguistics (ACL'2020)*.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 327–335, Sydney, Australia. Association for Computational Linguistics.

Ingmar Weber, Venkata R Kiran Garimella, and Alaa Batayneh. 2013. Secular vs. islamist polarization in Egypt on Twitter. In *ASONAM 2013*, pages 290–297.

Felix Ming Fai Wong, Chee Wei Tan, Soumya Sen, and Mung Chiang. 2013. Quantifying political leaning from tweets and retweets. In *ICWSM 2013*, pages 640–649.

Guido Zarrella and Amy Marsh. 2016. Mitre at semeval-2016 task 6: Transfer learning for stance detection. *arXiv preprint arXiv:1606.03784*.