Detecting Stance in Tweets And Analyzing its Interaction with Sentiment

Parinaz Sobhani
EECS, University of Ottawa
psobh090@uottawa.ca

Saif M. Mohammad and Svetlana Kiritchenko
National Research Council Canada
{saif.mohammad,svetlana.kiritchenko}@nrc-cnrc.gc.ca

Abstract

One may express favor (or disfavor) towards a target by using positive or negative language. Here for the first time we present a dataset of tweets annotated for whether the tweeter is in favor of or against pre-chosen targets, as well as for sentiment. These targets may or may not be referred to in the tweets, and they may or may not be the target of opinion in the tweets. We develop a simple stance detection system that outperforms all 19 teams that participated in a recent shared task competition on the same dataset (SemEval-2016 Task #6). Additionally, access to both stance and sentiment annotations allows us to conduct several experiments to tease out their interactions. We show that while sentiment features are useful for stance classification, they alone are not sufficient. We also show the impacts of various features on detecting stance and sentiment, respectively.

1 Introduction

Stance detection is the task of automatically determining from text whether the author of the text is in favor of, against, or neutral towards a proposition or target. The target may be a person, an organization, a government policy, a movement, a product, etc. For example, one can infer from Barack Obama's speeches that he is in favor of stricter gun laws in the US. Similarly, people often express stance towards various target entities through posts on online forums, blogs, Twitter, Youtube, Instagram, etc.

Automatically detecting stance has widespread applications in information retrieval, text summarization, and textual entailment. Over the last decade, there has been active research in modeling stance. However, most work focuses on congressional debates (Thomas et al., 2006) or debates in online forums (Somasundaran and Wiebe, 2009; Murakami and Raymond, 2010; Anand et al., 2011; Walker et al., 2012; Hasan and Ng, 2013). Here we explore the task of detecting stance in Twitter—a popular microblogging platform where people often express stance implicitly or explicitly.

The task we explore is formulated as follows: given a tweet text and a target entity (person, organization, movement, policy, etc.), automatic natural language systems must determine whether the tweeter is in favor of the given target, against the given target, or whether neither inference is likely. For example, consider the target–tweet pair:

(1)
Target: legalization of abortion
Tweet: *The pregnant are more than walking incubators, and have rights!*

Humans can deduce from the tweet that the tweeter is likely in favor of the target.[1]

Note that lack of evidence for 'favor' or 'against', does not imply that the tweeter is neutral towards the target. It may just mean that we cannot deduce stance from the tweet. In fact, this is a common phenomenon. On the other hand, the number of tweets from which we can infer neutral stance is expected to be small. An example is shown below:

(2)
Target: Hillary Clinton
Tweet: *Hillary Clinton has some strengths and some weaknesses.*

Stance detection is related to, but different from, sentiment analysis. Sentiment analysis tasks are

[1]Note that we use 'tweet' to refer to the text of the tweet and not to its meta-information. In our annotation task, we asked respondents to label for stance towards a given target based on the tweet text alone. However, automatic systems may benefit from exploiting tweet meta-information.

*Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics (*SEM 2016)*, pages 159–169,
Berlin, Germany, August 11-12, 2016.

formulated as determining whether a piece of text is positive, negative, or neutral, or determining from text the speaker's opinion and the target of the opinion (the entity towards which opinion is expressed). However, in stance detection, systems are to determine favorability towards a given (prechosen) target of interest. The target of interest may not be explicitly mentioned in the text and it may not be the target of opinion in the text. For example, consider the target–tweet pair below:

Target: *Donald Trump* (3)
Tweet: *Jeb Bush is the only sane candidate in this republican lineup.*

The target of opinion in the tweet is Jeb Bush, but the given target of interest is Donald Trump. Nonetheless, we can infer that the tweeter is likely to be unfavorable towards Donald Trump. Also note that, in stance detection, the target can be expressed in different ways which impacts whether the instance is labeled 'favor' or 'against'. For example, the target in example 1 could have been phrased as 'pro-life movement', in which case the correct label for that instance is 'against'. Also, the same stance (favor or against) towards a given target can be deduced from positive tweets and negative tweets. This interaction between sentiment and stance has not been adequately addressed in past work, and an important reason for this is the lack of a dataset annotated for both stance and sentiment.

Our contributions are as follows:

(1) We create the first tweets dataset labeled for stance, target of opinion, and sentiment. More than 4,000 tweets are annotated for whether one can deduce favorable or unfavorable stance towards one of five targets 'Atheism', 'Climate Change is a Real Concern', 'Feminist Movement', 'Hillary Clinton', and 'Legalization of Abortion'. Each of these tweets is also annotated for whether the target of opinion expressed in the tweet is the same as the given target of interest. Finally, each tweet is annotated for whether it conveys positive, negative, or neutral sentiment.

(2) Partitions of this stance-annotated data were used as training and test sets in the SemEval-2016 shared task competition 'Task #6: Detecting Stance in Tweets' (Mohammad et al., 2016b). Participants were provided with 2,914 training instances labeled for stance for the five targets. The test data included 1,249 instances. The task received submissions from 19 teams. The best performing system obtained an overall average F-score of 67.82. Their approach employed two recurrent neural network (RNN) classifiers: the first was trained to predict task-relevant hashtags on a very large unlabeled Twitter corpus. This network was used to initialize a second RNN classifier, which was trained with the provided training data.

(3) We propose a stance detection system that is much simpler than the SemEval-2016 Task #6 winning system (described above), and yet obtains an even better F-score of 70.32 on the shared task's test set. We use a linear-kernel SVM classifier that relies on features drawn from the training instances—such as word and character n-grams—as well as those obtained using external resources—such as sentiment features from lexicons and word-embedding features from additional unlabeled data.

(4) We conduct experiments to better understand the interaction between stance and sentiment and the factors influencing their interaction. We use the gold labels to determine the extent to which stance can be determined simply from sentiment. We apply the stance detection system (mentioned above in (3)), as a common text classification framework, to determine both stance and sentiment. Results show that while sentiment features are substantially useful for sentiment classification, they are not as effective for stance classification. Word embeddings improve the performance of both stance and sentiment classifiers. Further, even though both stance and sentiment detection are framed as three-way classification tasks on a common dataset where the majority class baselines are similar, automatic systems perform markedly better when detecting sentiment than when detecting stance towards a given target. Finally, we show that stance detection towards the target of interest is particularly challenging when the tweeter expresses opinion about an entity other than the target of interest. In fact, the text classification system performs close to majority baseline for such instances.

All of the stance data, including annotations for target of opinion and sentiment, are made freely available through the shared task website and the homepage for this Stance Project.[2]

[2]http://alt.qcri.org/semeval2016/task6/
www.saifmohammad.com/WebPages/StanceDataset.htm

Target	Example Favor Hashtag	Example Against Hashtag	Example Ambiguous Hashtag
Atheism	*#NoMoreReligions*	*#Godswill*	*#atheism*
Climate Change is Concern	-	*#globalwarminghoax*	*#climatechange*
Feminist Movement	*#INeedFeminismBecaus*	*#FeminismIsAwful*	*#Feminism*
Hillary Clinton	*#GOHILLARY*	*#WhyIAmNotVotingForHillary*	*#hillary2016*
Legalization of Abortion	*#proChoice*	*#prayToEndAbortion*	*#PlannedParenthood*

Table 1: Examples of stance-indicative and stance-ambiguous hashtags that were manually identified.

2 A Dataset for Stance from Tweets

The stance annotations we use are described in detail in Mohammad et al. (2016a). We summarize below how we compiled a set of tweets and targets for stance annotation, the questionnaire and crowdsourcing setup used for stance annotation, and an analysis of the stance annotations.

We first identified a list of target entities that were commonly known in the United States and also topics of debate: 'Atheism', 'Climate Change is a Real Concern", 'Feminist Movement', 'Hillary Clinton', and 'Legalization of Abortion'. Next, we compiled a small list of hashtags, which we will call *query hashtags*, that people use when tweeting about the targets. We split these hashtags into three categories: (1) *favor hashtags*: expected to occur in tweets expressing favorable stance towards the target (for example, *#Hillary4President*), (2) *against hashtags:* expected to occur in tweets expressing opposition to the target (for example, *#HillNo*), and (3) *stance-ambiguous hashtags:* expected to occur in tweets about the target, but are not explicitly indicative of stance (for example, *#Hillary2016*). Table 1 lists examples of hashtags used for each of the targets.

Next, we polled the Twitter API to collect close to 2 million tweets containing these hashtags (query hashtags). We discarded retweets and tweets with URLs. We kept only those tweets where the query hashtags appeared at the end. This reduced the number of tweets to about 1.7 million. We removed the query hashtags from the tweets to exclude obvious cues for the classification task. Since we only select tweets that have the query hashtag at the end, removing them from the tweet often still results in text that is understandable and grammatical.

Note that the presence of a stance-indicative hashtag is not a guarantee that the tweet will have the same stance.[3] Further, removal of query hashtags may result in a tweet that no longer expresses the same stance as with the query hashtag. Thus we manually annotate the tweet–target pairs after the pre-processing described above. For each target, we sampled an equal number of tweets pertaining to the favor hashtags, the against hashtags, and the stance-ambiguous hashtags. This helps in obtaining a sufficient number of tweets pertaining to each of the stance categories. Note that removing the query hashtag can sometimes result in tweets that do not explicitly mention the target. Consider:

> Target: Hillary Clinton (4)
> Tweet: *Benghazi must be answered for #Jeb16*

The query hashtags '#HillNo' was removed from the original tweet, leaving no mention of Hillary Clinton. Yet there is sufficient evidence (through references to Benghazi and #Jeb16) that the tweeter is likely against Hillary Clinton. Further, conceptual targets such as 'legalization of abortion' (much more so than person-name targets) have many instances where the target is not explicitly mentioned.

2.1 Stance Annotation

The core instructions given to annotators for determining stance are shown below.[4] Additional descriptions within each option (not shown here) make clear that stance can be expressed in many different ways, for example by explicitly supporting or opposing the target, by supporting an entity aligned with or opposed to the target, by re-tweeting somebody else's tweet, etc. We also asked a second question pertaining to whether the target of opinion expressed in the tweet is the same as the given target of interest.

[3] A tweet that has a seemingly favorable hashtag may in fact oppose the target; and this is not uncommon.

[4] The full set of instructions is made available on the shared task website: http://alt.qcri.org/semeval2016/task6/.

Target of Interest: [target entity]
Tweet: [tweet with query hashtag removed]

Q: From reading the tweet, which of the options below is most likely to be true about the tweeter's stance or outlook towards the target:

1. We can infer from the tweet that the tweeter supports the target
2. We can infer from the tweet that the tweeter is against the target
3. We can infer from the tweet that the tweeter has a neutral stance towards the target
4. There is no clue in the tweet to reveal the stance of the tweeter towards the target (support/against/neutral)

Q2: From reading the tweet, which of the options below is most likely to be true about the focus of opinion/sentiment in the tweet:

1. The tweet explicitly expresses opinion/sentiment about the target
2. The tweet expresses opinion/sentiment about something/someone other than the target
3. The tweet is not expressing opinion/sentiment

For each of the five selected targets, we randomly sampled 1,000 tweets from the 1.7 million tweets initially gathered from Twitter. Each of these tweets was uploaded on CrowdFlower for annotation as per the questionnaire shown above.[5] Each instance was annotated by at least eight annotators. For each target, the data not annotated for stance is used as the *domain corpus*—a set of unlabeled tweets that can be used to obtain information helpful to determine stance, such as relationships between relevant entities.

2.2 Analysis of Stance Annotations

The number of instances that were marked as neutral stance (option 3 in question 1) was less than 1%. Thus, we merged options 3 and 4 into one 'neither in favor nor against' option ('neither' for short). The inter-annotator agreement was 73.1% for question 1 (stance) and 66.2% for Question 2 (target of opinion).[6] These statistics are for the complete annotated dataset, which include instances that were genuinely difficult to annotate for stance (possibly because the tweets were too ungrammatical or vague) and/or instances that received poor annotations from the crowd workers (possibly because the particular annotator did not understand the tweet or its context). We selected instances with agreement equal or greater than 60% (at least 5 out of 8 annotators must agree) on Question 1 (stance) to create a dataset for machine learning experiments.[7] We will refer to this dataset as the *Stance Dataset*. The inter-annotator agreement on this Stance Dataset is 81.85% for question 1 (stance) and 68.9% for Question 2 (target of opinion). The rest of the instances are kept aside for future investigation. We partitioned the Stance Dataset into training and test sets based on the timestamps of the tweets. All annotated tweets were ordered by their timestamps, and the first 70% of the tweets formed the training set and the last 30% formed the test set. Table 2 shows the distribution of instances in the Stance Dataset.

Table 3 shows the distribution of responses to Question 2 (whether opinion is expressed directly about the given target). Observe that the percentage of 'opinion towards other' varies across different targets from 27% to 46%. Table 4 shows the distribution of instances by target of opinion for the 'favor' and 'against' stance labels. Observe that, as in Example 3, in a number of tweets from which we can infer unfavorable stance towards a target, the target of opinion is someone/something other than the target (about 26.5%). Manual inspection of the data also revealed that in a number of instances, the target is not directly mentioned, and yet stance towards the target was determined by the annotators. About 28% of the 'Hillary Clinton' instances and 67% of the 'Legalization of Abortion' instances were found to be of this kind—they did not mention 'Hillary' or 'Clinton' and did not mention 'abortion', 'pro-life', and 'pro-choice', respectively (case insensitive; with or without hashtag; with or without hyphen). Examples (1) and (4) shown earlier are instances of this, and are taken from our dataset.

3 Labeling the Stance Set for Sentiment

A key research question is the extent to which sentiment is correlated with stance. To that end, we annotated the same Stance Train and Test datasets described above for sentiment in a separate annotation project a few months later. We followed a procedure for annotation on CrowdFlower similar to that described above for stance, but now provided only the tweet (no target). We asked respondents to label the tweets as either positive, negative, or neither. The 'neither' category includes

[5] http://www.crowdflower.com

[6] We report absolute agreements here.

[7] The 60% threshold is somewhat arbitrary, but it seemed appropriate in terms of balancing confidence in the majority annotation and having to discard too many instances.

Target	# total	# train	% of instances in Train			# test	% of instances in Test		
			favor	against	neither		favor	against	neither
Atheism	733	513	17.9	59.3	22.8	220	14.5	72.7	12.7
Climate Change is Concern	564	395	53.7	3.8	42.5	169	72.8	6.5	20.7
Feminist Movement	949	664	31.6	49.4	19.0	285	20.4	64.2	15.4
Hillary Clinton	984	689	17.1	57.0	25.8	295	15.3	58.3	26.4
Legalization of Abortion	933	653	18.5	54.4	27.1	280	16.4	67.5	16.1
Total	4163	2914	25.8	47.9	26.3	1249	23.1	51.8	25.1

Table 2: Distribution of instances in the Stance Train and Test sets for Question 1 (Stance).

Target	Opinion towards		
	Target	Other	No one
Atheism	49.3	46.4	4.4
Climate Change is Concern	60.8	30.5	8.7
Feminist Movement	68.3	27.4	4.3
Hillary Clinton	60.3	35.1	4.6
Legalization of Abortion	63.7	31.0	5.4
Total	61.0	33.8	5.2

Table 3: Distribution of instances in the Stance dataset for Question 2 (Target of Opinion).

Stance	Opinion towards		
	Target	Other	No one
favor	94.2	5.1	0.7
against	72.8	26.5	0.7

Table 4: Distribution of target of opinion across stance labels.

mixed and neutral sentiment.

The inter-annotator agreement on the sentiment responses was 85.6%. Table 5 shows the distribution of sentiment labels in the training and test sets. Note that tweets corresponding to all targets, except 'Atheism', are predominantly negative.

4 A Common Text Classification Framework for Stance and Sentiment

Past work has shown that the most useful features for sentiment analysis are word and character n-grams and sentiment lexicons, whereas others such as negation features, part-of-speech features, and punctuation have a smaller impact (Wilson et al., 2013; Mohammad et al., 2013; Kiritchenko et al., 2014b; Rosenthal et al., 2015). More recently, features drawn from word embeddings have been shown to be effective in various text classification tasks such as sentiment analysis (Tang et al., 2014) and named entity recognition (Turian et al., 2010). All of these features are expected to be useful in stance classification as well. However, it is unclear which features will be more useful (and to what extent) for detecting stance as opposed to sentiment. Since we now have a dataset annotated for both stance and sentiment, we create a common text classification system (machine learning framework and features) and apply it to the Stance Dataset for detecting both stance and sentiment.

There is one exception to the common machine learning framework. The words and concepts used in tweets corresponding to the three stance categories are not expected to generalize across the targets. Thus, the stance system learns a separate model from training data pertaining to each of the targets.[8] Positive and negative language tend to have sufficient amount of commonality regardless of topic of discussion, and hence sentiment analysis systems traditionally learn a single model from all of the training data (Liu, 2015; Kiritchenko et al., 2014b; Rosenthal et al., 2015). Thus our sentiment experiments are also based on a single model trained on all of the Stance Training set.[9]

Tweets are tokenized and part-of-speech tagged with the CMU Twitter NLP tool (Gimpel et al., 2011). We train a linear-kernel Support Vector Machine (SVM) classifier on the Stance training set. SVM is a state-of-the-art learning algorithm proved to be effective on text categorization tasks and robust on large feature spaces. The SVM parameters are tuned using 5-fold cross-validation on Stance Training set. We used the implementation provided in Scikit-learn Machine Learning library (Pedregosa et al., 2011).

The features used in our text classification system are described below:[10]

- *n-grams*: presence or absence of contiguous sequences of 1, 2 and 3 tokens (word n-grams); presence or absence of contiguous sequences of 2, 3, 4, and 5 characters (character n-grams);
- *word embeddings*: the average of the word vectors for words appearing in a given

[8] We built a stance system that learns a single model from all training tweets, but its performance was worse.

[9] Training different models for each target did not yield better results.

[10] Use of tweet meta-information is left for future work.

Target	% of instances in Train			% of instances in Test		
	positive	negative	neither	positive	negative	neither
Atheism	60.4	35.1	4.5	59.1	35.5	5.5
Climate Change is Concern	31.7	49.6	18.7	29.6	51.5	18.9
Feminist Movement	17.9	77.3	4.8	19.3	76.1	4.6
Hillary Clinton	32.1	64.0	3.9	25.8	70.2	4.1
Legalization of Abortion	28.8	66.2	5.1	20.4	72.1	7.5
Total	33.1	60.5	6.5	29.5	63.3	7.2

Table 5: Distribution of sentiment in the Stance Train and Test sets.

tweet.[11] We derive 100-dimensional word vectors using Word2Vec Skip-gram model (Mikolov et al., 2013) trained over the Domain Corpus. (Recall that the Domain Corpus is the large set of unlabeled tweets pertaining to the five targets that were not manually labeled for stance).

- *sentiment features*: features drawn from sentiment lexicons as suggested in (Kiritchenko et al., 2014b). The lexicons used include NRC Emotion Lexicon (Mohammad and Turney, 2010), Hu and Liu Lexicon (Hu and Liu, 2004), MPQA Subjectivity Lexicon (Wilson et al., 2005), and NRC Hashtag Sentiment and Emoticon Lexicons (Kiritchenko et al., 2014b).

Some other feature sets that we experimented with, via cross-validation on the training set, included word embeddings trained on a generic Twitter corpus (not the domain corpus), the number of occurrences of each part-of-speech tag, the number of repeated sequences of exclamation or question marks, and the number of words with one character repeated more than two times (for example, *yessss*). However, they did not improve results there, and so we did not include them for the test set experiments.

We evaluate the learned models on the Stance Test set. As the evaluation measure, we use the average of the F1-scores (the harmonic mean of precision and recall) for the two main classes:[12]

For stance classification:

$$F_{avg} = \frac{F_{favor} + F_{against}}{2}$$

For sentiment classification:

$$F_{avg} = \frac{F_{positive} + F_{negative}}{2}$$

[11] Averaging is a commonly used vector combination method, although other approaches can also be pursued.

[12] A similar metric was used in the past for sentiment analysis—SemEval 2013 Task 2 (Wilson et al., 2013).

Note that F_{avg} can be determined for all of the test instances or for each target data separately. We will refer to the F_{avg} obtained through the former method as *F-micro-across-targets* or *F-microT* (for short). On the other hand, the F_{avg} obtained through the latter method, that is, by averaging the F_{avg} calculated for each target separately, will be called *F-macro-across-targets* or *F-macroT* (for short). F-microT was used as the bottom-line evaluation metric in the SemEval-2016 shared task on stance detection. Note that systems that perform relatively better on the more frequent target classes will obtain higher F-microT scores. On the other hand, to obtain a high F-macroT score a system has to perform well on all target classes.

5 Results of Automatic Systems

In the two subsections below, we present results obtained by the classifiers described above on detecting stance and sentiment, respectively, on the Stance Test set. (Cross-validation experiments on the Stance Training set produced similar results—and are thus not shown.)

5.1 Results for Stance Classification

Table 6 shows the overall results obtained by the automatic stance classifiers. Row i. shows results obtained by a random classifier (a classifier that randomly assigns a stance class to each instance), and row ii. shows results obtained by the majority classifier (a classifier that simply labels every instance with the majority class per target). Observe that F-microT for the majority class baseline is noticeably high. This is mostly due to the differences in the class distributions for the five targets: for most of the targets the majority of the instances are labeled as 'against' whereas for target 'Climate Change is a Real Concern' most of the data are labeled as 'favor'. Therefore, the F-scores for the classes 'favor' and 'against' are more balanced over all targets than for just

Classifier	F-macroT	F-microT
Benchmarks		
i. random	32.30	34.61
ii. majority	40.09	65.22
iii. first in SemEval'16 Task #6	56.03	67.82
iv. oracle sentiment	53.10	57.20
Our Classifiers		
a. n-grams	58.01	68.98
b. n-grams, embeddings	59.08	**70.32**
c. n-grams, sentiment lexicons	56.40	66.81
d. all three feature sets	**59.21**	69.84

Table 6: Stance Classification: Results obtained by automatic systems.

one target. Row iii. shows results obtained by the winning system (among nineteen participating teams) in the SemEval-2016 shared task on this data.

Results of an Oracle Sentiment Benchmark: The Stance Dataset with labels for both stance and sentiment allows us, for the first time, to conduct an experiment to determine the extent to which stance detection can be solved with sentiment analysis alone. Specifically, we determine the performance of an oracle system that assigns stance as follows: For each target, select a sentiment-to-stance assignment (mapping all positive instances to 'favor' and all negative instances to 'against' OR mapping all positive instances to 'against' and all negative instances to 'favor') that maximizes the F-macroT score. We call this benchmark the Oracle Sentiment Benchmark. This benchmark is informative because it gives an upper bound of the F-score one can expect when using a traditional sentiment system for stance detection by simply mapping sentiment labels to stance labels.

Row iv. in Table 6 shows the F-scores obtained by the Oracle Sentiment Benchmark on the test set. Observe that the F-macroT score is markedly higher than the corresponding score for the majority baseline, but yet much lower than 100%. This shows that even though sentiment can play a key role in detecting stance, sentiment alone is not sufficient.

Results Obtained by Our Classifier: Rows a., b., c., and d. show results obtained by our SVM classifier using n-gram features alone, n-grams and word embedding features, n-grams and sentiment lexicon features, and n-grams, word embeddings, and sentiment lexicon features ('all three feature sets'), respectively. The results in row a. show the performance that can be achieved on this test set using only the provided training data and no external resources (such as lexicons and extra labeled or unlabeled data). Observe that the results obtained by our system surpass the results obtained by the winning team in the SemEval shared task (row iii.). Also note that while the n-grams and word embeddings alone provide the highest F-microT score, the sentiment lexicon features are beneficial if one is interested in a higher F-macroT score. Table 7 shows F-scores for tweets pertaining to each of the targets. Observe that the word embedding features are beneficial for four out of five targets. The sentiment lexicon features bring additional improvements for two targets, 'Atheism' and 'Hillary Clinton'.

Recall that the Stance Dataset is also annotated for whether opinion is expressed directly about the target, about somebody/someone other than the target, or no opinion is being expressed. Table 8 shows stance detection F-scores obtained on tweets that express opinion directly towards the target and on tweets that express opinion towards others. (The number of tweets for 'no opinion is being expressed' is small, and thus not covered in this experiment.) Observe that the performance of the classifier is considerably better for tweets where opinion is expressed towards the target, than otherwise. Detecting stance towards a given target from tweets that express opinion about some other entity has not been addressed in our research community, and results in Table 8 show that it is particularly challenging. We hope that this dataset will encourage more work to address this gap in performance.

5.2 Results for Sentiment Classification

Table 9 shows results obtained by various automatic classification systems on the sentiment labels of the Stance Dataset. Observe that the scores obtained by the majority class baseline for the three-way sentiment classification is similar to the majority class baseline for the three-way stance classification. Nonetheless, the text classification system obtains markedly higher scores on sentiment prediction than on predicting stance. Observe also that on this sentiment task (unlike the stance task) the sentiment lexicon features are particularly useful (see row b.). Word embeddings features provide improvements over n-grams (row c.); however, adding them on top of n-grams and

Classifier	Atheism	Climate Concern	Feminist Movement	Hillary Clinton	Legalization of Abortion	F-macroT	F-microT
Majority classifier	42.11	42.12	39.10	36.83	40.30	40.09	65.22
Our classifiers							
a. n-grams	65.19	42.35	57.46	58.63	66.42	58.01	68.98
b. n-grams, embeddings	68.25	**43.80**	**58.72**	57.74	**66.91**	59.08	**70.32**
c. n-grams, sentiment lexicons	65.17	40.08	54.48	60.56	61.70	56.40	66.81
d. all three feature sets	**69.19**	42.35	56.11	**61.74**	66.70	**59.21**	69.84

Table 7: Stance Classification: F-scores obtained for each of the targets (the columns) when one or more of the feature groups are added. Highest scores in each column is shown in bold.

	F-macroT		**F-microT**	
Classifier	Target	Other	Target	Other
all three features	63.51	38.14	75.31	44.15

Table 8: Stance Classification: F-scores obtained for tweets with opinion towards the target and tweets with opinion towards another entity.

Classifier	F_{Pos}	F_{Neg}	F-microT
Majority classifier	44.22	78.35	61.28
Our classifiers			
a. n-grams	64.78	81.75	73.27
b. n-grams, sentiment lex.	**72.21**	**85.52**	**78.87**
c. n-grams, embeddings	68.85	84.00	76.43
d. all three feature sets	71.90	85.21	78.56

Table 9: Sentiment Classification: Results obtained by automatic systems.

	Opinion towards	
Classifier	Target	Other
all three feature sets	79.64	77.81

Table 10: Sentiment Classification: F-microT on tweets with opinion towards the target and tweets with opinion towards another entity.

sentiment features is not beneficial (row d.).

Table 10 shows the performance of the sentiment classifier on tweets that express opinion towards the given target and those that express opinion about another entity. Observe that the sentiment prediction performance (unlike stance prediction performance) is similar on the two sets of tweets. This shows that the two sets of tweets are not qualitatively different in how they express opinion. However, since one set expresses opinion about an entity other than the target of interest, detecting stance towards the target of interest from them is notably more challenging.

6 Related Work

SemEval-2016 Task #6. The SemEval-2016 Task 'Detecting Stance in Tweets' received submissions from 19 teams, wherein the highest classification F-score obtained was 67.82. The best performing systems used standard text classification features such as those drawn from n-grams, word vectors, and sentiment lexicons. Some teams drew additional gains from noisy stance-labeled data created using distant supervision techniques. A large number of teams used word embeddings and some used deep neural networks such as RNNs and convolutional neural nets. Nonetheless, none of these systems surpassed our results presented here.

Other Stance Detection Work. In work by Somasundaran and Wiebe (2010), a lexicon for detecting argument trigger expressions was created and subsequently leveraged to identify arguments. These extracted arguments, together with sentiment expressions and their targets, were employed in a supervised learner as features for stance classification. Anand et al. (2011) deployed a rule-based classifier with several features such as unigrams, bigrams, punctuation marks, syntactic dependencies and the dialogic structure of the posts. Here, we did not explore dependency features since dependency parsers are not as accurate on tweets. Additionally, Anand et al. (2011) showed that there is no significant difference in performance between systems that use only unigrams and systems that also use other features such as LIWC and opinion or POS generalized dependencies in stance classification. Some of these features were used by the teams participating in the SemEval task over this dataset; however, their systems' performances were lower than the performance showed by our stance detection system. The dialogic relations of agreements and disagreements between posts were exploited by Walker et al. (2012). These relationships are not provided for our Stance dataset.

Sobhani et al. (2015) extracted arguments used in online news comments to leverage them as extra features for detecting stance. Faulkner (2014) investigated the problem of detecting document-level stance in student essays by making use of

two sets of features that are supposed to represent stance-taking language. Deng and Wiebe (2014) investigated the relationships and interactions among entities and events explicitly mentioned in the text with the goal of improving sentiment classification. In stance classification, however, the predetermined target of interest may not be mentioned in the text, or may not be the target of opinion in the text. Rajadesingan and Liu (2014) determined stance at user level based on the assumption that if several users retweet one pair of tweets about a controversial topic, it is likely that they support the same side of a debate. In this work, we focus on detecting stance, as well as possible, from a single tweet. Features that help to this end will likely also be useful when there is access to multiple tweets from the same tweeter.

Sentiment Analysis and Related Tasks. There is a vast amount of work in sentiment analysis of tweets, and we refer the reader to surveys (Pang and Lee, 2008; Liu and Zhang, 2012; Mohammad, 2015) and proceedings of recent shared task competitions (Wilson et al., 2013; Rosenthal et al., 2015). Closely-related is the area of aspect based sentiment analysis (ABSA), where the goal is to determine sentiment towards aspects of a product such as speed of processor and screen resolution of a cell phone. We refer the reader to SemEval proceedings for related work on ABSA (Pontiki et al., 2015; Pontiki et al., 2014). Mohammad et al. (2013) and Kiritchenko et al. (2014a) came first in the SemEval-2013 Sentiment in Twitter and SemEval-2014 ABSA shared tasks. We use most of the features they proposed in our classifier. There are other subtasks in opinion mining related to stance classification, such as biased language detection (Recasens et al., 2013; Yano et al., 2010), perspective identification (Lin et al., 2006) and user classification based on their views (Kato et al., 2008). Perspective identification was defined as the subjective evaluation of points of view (Lin et al., 2006). None of the prior work has created a dataset annotated for both stance and sentiment.

7 Conclusions and Future Work

We presented the first dataset of tweets annotated for both stance towards given targets and sentiment. Partitions of the stance-annotated data created as part of this project were used as training and test sets in the SemEval-2016 shared task 'Task #6: Detecting Stance in Tweets' that received submissions from 19 teams. We proposed a simple, but effective stance detection system that obtained an F-score (70.32) higher than the one obtained by the more complex, best-performing system in the competition. We used a linear-kernel SVM classifier that leveraged word and character n-grams as well as sentiment features drawn from available sentiment lexicons and word-embedding features drawn from additional unlabeled data.

Finally, we conducted several experiments to tease out the interactions between the stance and sentiment. Notably, we showed that even though sentiment features are useful for stance detection, they alone are not sufficient. We also showed that even though humans are capable of detecting stance towards a given target from texts that express opinion towards a different target, automatic systems perform poorly on such data.

The features we used are not new to the community and not specifically tailored for stance detection. Nonetheless, they outperform those developed by the 19 teams that participated in the SemEval-2016 shared task on this dataset. This emphasizes the need for more research in exploring novel techniques specifically suited for detecting stance. Some avenues of future work include obtaining more sophisticated features such as those derived from dependency parse trees and automatically generated entity–entity relationship knowledge bases. Knowing that entity X is an adversary of entity Y can be useful in detecting stance towards Y from tweets that express opinion about X. One may also pursue more sophisticated classifiers, for example, deep architectures that jointly model stance, target of opinion, and sentiment. We are also interested in developing stance detection systems that do not require stance-labeled instances for the target of interest, but instead, can learn from existing stance-labeled instances for other targets in the same domain. We also want to model the ways in which stance is conveyed, and how the distribution of stance towards a target changes over time.

Acknowledgments

We thank Colin Cherry and Xiaodan Zhu for helpful discussions. The first author of this paper was supported by the Natural Sciences and Engineering Research Council of Canada under the CREATE program.

References

Pranav Anand, Marilyn Walker, Rob Abbott, Jean E. Fox Tree, Robeson Bowmani, and Michael Minor. 2011. Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pages 1–9.

Lingjia Deng and Janyce Wiebe. 2014. Sentiment propagation via implicature constraints. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pages 377–385, Sweden.

Adam Faulkner. 2014. Automated classification of stance in student essays: An approach using stance target information and the Wikipedia link-based measure. In *Proceedings of the Flairs Conference*.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Kazi Saidul Hasan and Vincent Ng. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the International Joint Conference on Natural Language Processing*, pages 1348–1356.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177.

Yoshikiyo Kato, Sadao Kurohashi, Kentaro Inui, Robert Malouf, and Tony Mullen. 2008. Taking sides: User classification for informal online political discourse. *Internet Research*, 18(2):177–190.

Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif M. Mohammad. 2014a. NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the International Workshop on Semantic Evaluation*, Dublin, Ireland, August.

Svetlana Kiritchenko, Xiaodan Zhu, and Saif M. Mohammad. 2014b. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762.

Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann. 2006. Which side are you on? Identifying perspectives at the document and sentence levels. In *Proceedings of the Conference on Computational Natural Language Learning*, pages 109–116.

Bing Liu and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 415–463. Springer.

Bing Liu. 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Saif M. Mohammad and Peter D. Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34.

Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the International Workshop on Semantic Evaluation*, Atlanta, Georgia, USA, June.

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016a. A dataset for detecting stance in tweets. In *Proceedings of the Language Resources and Evaluation Conference*, Portorož, Slovenia.

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016b. SemEval-2016 Task 6: Detecting stance in tweets. In *Proceedings of the International Workshop on Semantic Evaluation*, San Diego, California.

Saif M Mohammad. 2015. Sentiment analysis: Detecting valence, emotions, and other affectual states from text.

Akiko Murakami and Rudy Raymond. 2010. Support or oppose? Classifying positions in online debates from reply activities and opinion expressions. In *Proceedings of the International Conference on Computational Linguistics*, pages 869–875, Beijing, China.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

Fabian Pedregosa, Gaël Varoquaux, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Maria Pontiki, Dimitrios Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 Task 4: Aspect based sentiment analysis. In *Proceedings of the International Workshop on Semantic Evaluation*, Dublin, Ireland, August.

Maria Pontiki, Dimitrios Galanis, Haris Papageogiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 Task 12: Aspect based sentiment analysis. In *Proceedings of the International Workshop on Semantic Evaluation*, Denver, Colorado.

Ashwin Rajadesingan and Huan Liu. 2014. Identifying users with opposing opinions in Twitter debates. In *Proceedings of the Conference on Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 153–160. Washington, DC, USA.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1650–1659.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif M. Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. SemEval-2015 Task 10: Sentiment analysis in Twitter. In *Proceedings of the International Workshop on Semantic Evaluations*.

Parinaz Sobhani, Diana Inkpen, and Stan Matwin. 2015. From argumentation mining to stance classification. In *Proceedings of the Workshop on Argumentation Mining*, pages 67–77, Denver, Colorado, USA.

Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 226–234, Suntec, Singapore.

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop CAAGET*, pages 116–124.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for Twitter sentiment classification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1555–1565.

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 Conference on Empirical Methods in Natural Language Processing*, pages 327–335.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 384–394.

Marilyn A. Walker, Pranav Anand, Robert Abbott, and Ricky Grant. 2012. Stance classification using dialogic properties of persuasion. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 592–596.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver, British Columbia, Canada.

Theresa Wilson, Zornitsa Kozareva, Preslav Nakov, Sara Rosenthal, Veselin Stoyanov, and Alan Ritter. 2013. SemEval-2013 Task 2: Sentiment analysis in Twitter. In *Proceedings of the International Workshop on Semantic Evaluation*, Atlanta, USA, June.

Tae Yano, Philip Resnik, and Noah A Smith. 2010. Shedding (a thousand points of) light on biased language. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 152–158.