SemEval-2016 Task 6: Detecting Stance in Tweets

Saif M. Mohammad
National Research Council Canada
saif.mohammad@nrc-cnrc.gc.ca

Svetlana Kiritchenko
National Research Council Canada
svetlana.kiritchenko@nrc-cnrc.gc.ca

Parinaz Sobhani
University of Ottawa
psobh090@uottawa.ca

Xiaodan Zhu
National Research Council Canada
xiaodan.zhu@nrc-cnrc.gc.ca

Colin Cherry
National Research Council Canada
colin.cherry@nrc-cnrc.gc.ca

Abstract

Here for the first time we present a shared task on detecting stance from tweets: given a tweet and a target entity (person, organization, 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. The target of interest may or may not be referred to in the tweet, and it may or may not be the target of opinion. Two tasks are proposed. Task A is a traditional supervised classification task where 70% of the annotated data for a target is used as training and the rest for testing. For Task B, we use as test data all of the instances for a new target (not used in task A) and no training data is provided. Our shared task received submissions from 19 teams for Task A and from 9 teams for Task B. The highest classification F-score obtained was 67.82 for Task A and 56.28 for Task B. However, systems found it markedly more difficult to infer stance towards the target of interest from tweets that express opinion towards another entity.

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.

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:

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

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

We annotated 4870 English tweets for stance towards six commonly known targets in the United States. The data corresponding to five of the targets ('Atheism', 'Climate Change is a Real Concern', 'Feminist Movement', 'Hillary Clinton', and 'Legalization of Abortion') was used in a standard supervised stance detection task – *Task A*. About 70% of the tweets per target were used for training and the remaining for testing. All of the data corresponding to the target 'Donald Trump' was used as test set in a separate task – *Task B*. No training data labeled with stance towards 'Donald Trump' was provided. However, participants were free to use data from Task A to develop their models for Task B.

[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 SemEval-2016, pages 31–41,
San Diego, California, June 16-17, 2016. ©2016 Association for Computational Linguistics

Task A received submissions from 19 teams, wherein the highest classification F-score obtained was 67.82. Task B, which is particularly challenging due to lack of training data, received submissions from 9 teams wherein the highest classification F-score obtained was 56.28. 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, for Task A, none of these systems surpassed a baseline SVM classifier that uses word and character n-grams as features (Mohammad et al., 2016b). Further, results are markedly worse for instances where the target of interest is not the target of opinion.

More gains can be expected in the future on both tasks, as researchers better understand this new task and data. All of the data, an interactive visualization of the data, and the evaluation scripts are available on the task website as well as the homepage for this Stance project.[2]

2 Subtleties of Stance Detection

In the sub-sections below we discuss some of the nuances of stance detection, including a discussion on neutral stance and the relationship between stance and sentiment.

2.1 Neutral Stance

The classification task formulated here does not include an explicit neutral class. The lack of evidence for 'favor' or 'against' does not imply that the tweeter is neutral towards the target. It may just be that one cannot deduce stance from the tweet. In fact, this is fairly common. 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:

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

Thus, even though we obtain annotations for neutral stance, we eventually merge all classes other than 'favor' and 'against' into one 'neither' class.

2.2 Stance and Sentiment

Stance detection is related to, but different from, sentiment analysis. Sentiment analysis tasks are usually 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 favour 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 (favour or against) towards a given target can be deduced from positive tweets and negative tweets. See Mohammad et al. (2016b) for a quantitative exploration of this interaction between stance and sentiment.

3 A Dataset for Stance from Tweets

The stance annotations we use are described in detail in Mohammad et al. (2016a). The same dataset was subsequently also annotated for target of opinion and sentiment (in addition to stance towards a given target) (Mohammad et al., 2016b). These additional annotations are not part of the SemEval-2016 competition, but are made available for future research. We summarize below all relevant details for this shared task: how we compiled a set of tweets and targets for stance annotation (Section 3.1), the questionnaire and crowdsourcing setup used for stance annotation (Section 3.2), and an analysis of the stance annotations (Section 3.3).

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

3.1 Selecting the Tweet–Target Pairs

We wanted to create a dataset of stance-labeled tweet–target pairs with the following properties:

1: The tweet and target are commonly understood by a wide number of people in the US. (The data was also eventually annotated for stance by respondents living in the US.)

2: There must be a significant amount of data for the three classes: favor, against, and neither.

3: Apart from tweets that explicitly mention the target, the dataset should include a significant number of tweets that express opinion towards the target without referring to it by name.

4: Apart from tweets that express opinion towards the target, the dataset should include a significant number of tweets in which the target of opinion is different from the given target of interest. Downstream applications often require stance towards particular pre-chosen targets of interest (for example, a company might be interested in stance towards its product). Having data where the target of opinion is some other entity (for example, a competitor's product) helps test how well stance detection systems can cope with such instances.

To help with Property 1, the authors of this paper compiled a list of target entities commonly known in the United States. (See Table 1 for the list.)

We created 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*). Next, we polled the Twitter API to collect over two million tweets containing these query hashtags. We discarded retweets and tweets with URLs. We kept only those tweets where the query hashtags appeared at the end. 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—up to 1000 tweets at most per target. This helps in obtaining a sufficient number of tweets pertaining to each of the stance categories (Property 2). Properties 3 and 4 are addressed to some extent by the fact 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.

3.2 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.

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:

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

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

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)

Each of the tweet–target pairs selected for annotation was uploaded on CrowdFlower for annotation with the questionnaire shown above.[5] Each instance was annotated by at least eight respondents.

3.3 Analysis of Stance Annotations

The number of instances that were marked as neutral stance (option 3) 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%. 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 to or greater than 60% (at least 5 out of 8 annotators must agree) to create the test and training sets for this task.[6] We will refer to this dataset as the *Stance Dataset*. The inter-annotator agreement on this set is 81.85%. 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 1 shows the number and distribution of instances in the Stance Dataset.

Inspection of the data revealed that often the target is not directly mentioned, and yet stance towards the target was determined by the annotators. About 30% of the 'Hillary Clinton' instances and 65% 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.

An interactive visualization of the Stance Dataset that shows various statistics about the data is available at the task website. Note that it also shows sentiment and target of opinion annotations (in addition to stance). Clicking on various visualization elements filters the data. For example, clicking on 'Feminism' and 'Favor' will show information pertaining to only those tweets that express favor towards feminism. One can also use the check boxes on the left to view only test or training data, or data on particular targets.

4 Task Setup: Automatic Stance Classification

The Stance Dataset was partitioned so as to be used in two tasks described in the subsections below: Task A (supervised framework) and Task B (weakly supervised framework). Participants could provide submissions for either one of the tasks, or both tasks. Both tasks required classification of tweet–target pairs into exactly one of three classes:

- Favor: We can infer from the tweet that the tweeter supports the target (e.g., directly or indirectly by supporting someone/something, by opposing or criticizing someone/something opposed to the target, or by echoing the stance of somebody else).
- Against: We can infer from the tweet that the tweeter is against the target (e.g., directly or indirectly by opposing or criticizing someone/something, by supporting someone/something opposed to the target, or by echoing the stance of somebody else).
- Neither: none of the above.

4.1 Task A: Supervised Framework

This task tested stance towards five targets:'Atheism', 'Climate Change is a Real Concern', 'Feminist Movement', 'Hillary Clinton', and 'Legalization of Abortion'. Participants were provided

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

[6] The 60% threshold is somewhat arbitrary, but it seemed appropriate in terms of balancing quality and quantity.

			% of instances in Train				% of instances in Test		
Target	**# total**	**# train**	favor	against	neither	**# test**	favor	against	neither
Data for Task A									
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
All	4163	2914	25.8	47.9	26.3	1249	24.3	57.3	18.4
Data for Task B									
Donald Trump	707	0	-	-	-	707	20.93	42.29	36.78

Table 1: Distribution of instances in the Stance Train and Test sets for Task A and Task B.

with 2,914 labeled training data instances for the five targets. The test data included 1,249 instances.

4.2 Task B: Weakly Supervised Framework

This task tested stance towards one target 'Donald Trump' in 707 tweets. Participants were not provided with any training data for this target. They were given about 78,000 tweets associated with 'Donald Trump' to various degrees – the *domain corpus*, but these tweets were not labeled for stance. These tweets were gathered by polling Twitter for hashtags associated with Donald Trump.

4.3 Common Evaluation Metric for Both Task A and Task B

We used the macro-average of the F1-score for 'favor' and the F1-score for 'against' as the bottom-line evaluation metric.

$$F_{avg} = \frac{F_{favor} + F_{against}}{2} \qquad (1)$$

where F_{favor} and $F_{against}$ are calculated as shown below:

$$F_{favor} = \frac{2P_{favor}R_{favor}}{P_{favor}+R_{favor}} \qquad (2)$$

$$F_{against} = \frac{2P_{against}R_{against}}{P_{against}+R_{against}} \qquad (3)$$

Note that the evaluation measure does not disregard the 'neither' class. By taking the average F-score for only the 'favor' and 'against' classes, we treat 'neither' as a class that is not of interest—or 'negative' class in Information Retrieval (IR) terms. Falsely labeling negative class instances still adversely affects the scores of this metric. If one uses simple accuracy as the evaluation metric, and if the negative class is very dominant (as is the case in IR), then simply labeling every instance with the negative class will obtain very high scores.

If one randomly accesses tweets, then the probability that one can infer 'favor' or 'against' stance towards a pre-chosen target of interest is small. This has motivated the IR-like metric used in this competition, even though we worked hard to have marked amounts of 'favor' and 'against' data in our training and test sets. This metric is also similar to how sentiment prediction was evaluated in recent SemEval competitions.

This evaluation metric can be seen as a micro-average of F-scores across targets (F-microT). Alternatively, one could determine the mean of the F_{avg} scores for each of the targets—the macro average across targets (F-macroT). Even though not the official competition metric, the F-macroT can easily be determined from the per-target F_{avg} scores shown in the result tables of Section 5.

The participants were provided with an evaluation script so that they could check the format of their submission and determine performance when gold labels were available.

5 Systems and Results for Task A

We now discuss various baseline systems and the official submissions to Task A.

5.1 Task A Baselines

Table 2 presents the results obtained with several baseline classifiers first presented in (Mohammad et al., 2016b). Since the baseline system was developed by some of the organizers of this task, it was

		Overall		Atheism	Climate	Feminism	Hillary	Abortion
Team	F_{favour}	$F_{against}$	F_{avg}	F_{avg}	F_{avg}	F_{avg}	F_{avg}	F_{avg}
Baselines								
Majority class	52.01	**78.44**	65.22	42.11	42.12	39.10	36.83	40.30
SVM-unigrams	54.49	72.13	63.31	53.25	38.39	55.65	57.02	60.09
SVM-ngrams	**62.98**	74.98	**68.98**	**65.19**	42.35	**57.46**	**58.63**	**66.42**
SVM-ngrams-comb	54.11	70.01	62.06	53.27	**47.76**	52.82	56.50	63.71
Participating Teams								
MITRE	59.32	**76.33**	**67.82**	61.47	41.63	**62.09**	57.67	57.28
pkudblab	**61.98**	72.67	67.33	63.34	52.69	51.33	64.41	61.09
TakeLab	60.93	72.73	66.83	**67.25**	41.25	53.01	**67.12**	61.38
PKULCWM	56.96	74.55	65.76	56.39	40.39	51.32	62.26	61.56
ECNU	60.55	70.54	65.55	61.97	41.32	56.21	57.85	61.25
CU-GWU	54.99	72.21	63.60	55.68	39.41	53.88	51.19	59.38
IUCL-RF	52.61	74.59	63.60	57.93	39.06	51.06	49.84	57.61
DeepStance	58.44	68.65	63.54	52.90	40.40	52.34	55.35	**63.32**
UWB	57.41	69.42	63.42	57.88	46.90	51.82	59.82	61.98
IDI@NTNU	58.97	65.97	62.47	59.59	**54.86**	48.59	57.89	54.47
Tohoku	49.25	75.18	62.21	58.90	39.51	52.41	39.81	37.75
ltl.uni-due	48.71	74.75	61.73	52.47	35.50	55.12	44.23	57.25
LitisMind	50.67	72.20	61.44	52.36	39.15	57.16	42.08	45.88
JU_NLP	46.68	74.53	60.60	38.99	42.60	45.65	50.25	41.83
NEUSA	49.03	71.20	60.12	48.90	41.95	52.14	48.53	61.89
nldsucsc	50.90	67.81	59.36	57.19	42.10	48.97	57.27	61.66
WFU/TNT	47.55	70.89	59.22	46.16	42.07	47.91	45.88	45.34
INESC-ID	50.58	64.57	57.58	52.67	44.92	49.00	50.64	49.93
Thomson Reuters	30.16	62.23	46.19	44.79	35.86	39.37	34.98	38.89

Table 2: Results for Task A, reporting the official competition metric as 'Overall F_{avg}', along with F_{favor} and $F_{against}$ over all targets and F_{avg} for each individual target. The highest scores in each column among the baselines and among the participating systems are shown in bold.

not entered as part of the official competition.
Baselines:

1. *Majority class*: a classifier that simply labels every instance with the majority class ('favor' or 'against') for the corresponding target;
2. *SVM-unigrams*: five SVM classifiers (one per target) trained on the corresponding training set for the target using word unigram features;
3. *SVM-ngrams*: five SVM classifiers (one per target) trained on the corresponding training set for the target using word n-grams (1-, 2-, and 3-gram) and character n-grams (2-, 3-, 4-, and 5-gram) features;
4. *SVM-ngrams-comb*: one SVM classifier trained on the combined (all 5 targets) training set using word n-grams (1-, 2-, and 3-gram) and character n-grams (2-, 3-, 4-, and 5-gram) features.

The SVM parameters were tuned using 5-fold cross-validation on the training data. The first three columns of the table show the official competition metric (Overall F_{avg}) along with the two components that are averaged to obtain it (F_{favor} and $F_{against}$). The next five columns describe per-target results—the official metric as calculated over each of the targets individually.

Observe that the Overall F_{avg} for the Majority class baseline is very 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 one target.

We can see that a supervised classifier using unigram features alone produces results markedly

	Opinion Towards		**All**
Team	Target	Other	
Baselines			
Majority class	71.27	41.33	65.22
SVM-unigrams	69.39	38.96	63.31
SVM-ngrams	**74.54**	**43.20**	**68.98**
SVM-ngrams-comb	66.60	38.05	62.06
Participating Teams			
MITRE	72.49	44.48	**67.82**
pkudblab	71.07	**46.66**	67.33
TakeLab	**73.66**	37.47	66.83
PKULCWM	70.62	45.89	65.76
ECNU	70.29	44.25	65.55
CU-GWU	67.89	45.28	63.60
IUCL-RF	67.77	41.96	63.60
DeepStance	67.81	44.00	63.54
UWB	67.60	44.54	63.42
IDI@NTNU	66.25	42.26	62.47
Tohoku	66.44	44.09	62.21
ltl.uni-due	67.23	42.45	61.73
LitisMind	66.42	41.27	61.44
JU_NLP	62.55	49.34	60.60
NEUSA	65.39	39.48	60.12
nldsucsc	65.71	34.64	59.36
WFU/TNT	67.28	34.89	59.22
INESC-ID	63.99	36.63	57.58
Thomson Reuters	49.98	32.43	46.19

Table 3: Results for Task A (the official competition metric F_{avg}) on different subsets of the test data. The highest scores in each column among the baselines and among the participating systems are shown in bold.

above the majority baseline for most of the targets. Furthermore, employing higher-order n-gram features results in substantial improvements for all targets as well as for the Overall F_{avg}. Training separate classifiers for each target seems a better solution than training a single classifier for all targets even though the combined classifier has access to significantly more data. As expected, the words and concepts used in tweets corresponding to the stance categories do not generalize well across the targets. However, there is one exception: the results for 'Climate Change' improve by over 5% when the combined classifier has access to the training data for other targets. This is probably because it has access to more balanced dataset and more representative instances for 'against' class. Most teams chose to train separate classifiers for different targets.

5.2 Task A Participating Stance Systems

Nineteen teams competed in Task A on supervised stance detection. Table 2 shows each team's performance, both in aggregate and in terms of individual targets. Teams are sorted in terms of their performance according to the official metric. The best results obtained by a participating system was an Overall F_{avg} of 67.82 by *MITRE*. 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 Task A data. However, this result is not higher than the SVM-ngrams baseline.

In general, per-target results are lower than the Overall F_{avg}. This is likely due to the fact that it is easier to balance 'favor' and 'against' classes over all targets than it is for exactly one target. That is, when dealing with all targets, one can use the natural abundance of tweets in favor of concern over climate change to balance against the fact that many of the other targets have a high proportion of tweets against them. Most systems were optimized for the competition metric, which allows cross-target balancing, and thus would naturally perform worse on per-target metrics. *IDI@NTNU* is an interesting exception, as their submission focused on the 'Climate Change' target, and they did succeed in producing the best result for that target.

We also calculated Task A results on two subsets of the test set: (1) a subset where opinion is expressed towards the target, (2) a subset where opinion is expressed towards some other entity. Table 3 shows these results. It also shows results on the complete test set (All), for easy reference. Observe that the stance task is markedly more difficult when stance is to be inferred from a tweet expressing opinion about some other entity (and not the target of interest). This is not surprising because it is a more challenging task, and because there has been very little work on this in the past.

5.3 Discussion

Most teams used standard text classification features such as n-grams and word embedding vectors, as well as standard sentiment analysis features such as those drawn from sentiment lexicons (Kiritchenko et al., 2014b). Some teams polled Twitter for stance-bearing hashtags, creating additional noisy stance data. Three teams tried variants of this strategy: *MITRE*, *DeepStance* and *nldsucsc*. These teams are distributed somewhat evenly throughout the standings, and although *MITRE* did use extra data in its top-placing entry, *pkudblab* achieved nearly the same score with only the provided data.

Another possible differentiator would be the use of continuous word representations, derived either from extremely large sources such as Google News, directly from Twitter corpora, or as a by-product of training a neural network classifier. Nine of the nineteen entries used some form of word embedding, including the top three entries, but *PKULCWM*'s fourth place result shows that it is possible to do well with a more traditional approach that relies instead on Twitter-specific linguistic pre-processing. Along these lines, it is worth noting that both *MITRE* and *pkudblab* reflect knowledge-light approaches to the problem, each relying minimally on linguistic processing and external lexicons.

Seven of the nineteen submissions made extensive use of publicly-available sentiment and emotion lexicons such as the 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 Lexicons (Kiritchenko et al., 2014b).

Recall that the SVM-ngrams baseline also performed very well, using only word and character n-grams in its classifiers. This helps emphasize the fact that for this young task, the community is still a long way from an established set of best practices.

6 Systems and Results for Task B

The sub-sections below discuss baselines and official submissions to Task B. Recall, that the test data for Task B is for the target 'Donald Trump', and no training data for this target was provided.

6.1 Task B Baselines

We calculated two baselines listed below:

1. *Majority class*: a classifier that simply labels every instance with the majority class ('favor' or 'against') for the corresponding target;
2. *SVM-ngrams-comb*: one SVM classifier trained on the combined (all 5 targets) Task A training set, using word n-grams (1-, 2-, and 3-gram) and character n-grams (2-, 3-, 4-, and 5-gram) features.

Team	F_{favor}	$F_{against}$	F_{avg}
Baselines			
Majority class	0.00	59.44	29.72
SVM-ngrams-comb	18.42	38.45	28.43
Participating Teams			
pkudblab	**57.39**	55.17	**56.28**
LitisMind	30.04	**59.28**	44.66
INF-UFRGS	32.56	52.09	42.32
UWB	34.26	49.78	42.02
ECNU	17.96	50.20	34.08
USFD	10.93	54.46	32.70
Thomson Reuters	14.39	50.39	32.39
ltl.uni-due	46.56	05.71	26.14
NEUSA	16.59	34.87	25.73

Table 4: Results for Task B, reporting the official competition metric as F_{avg}, along with F_{favor} and $F_{against}$. The highest score in each column is shown in bold.

The results are presented in Table 4. Note that the class distribution for the target 'Donald Trump' is more balanced. Therefore, the F_{avg} for the Majority baseline for this target is much lower than the corresponding values for other targets. Yet, the combined classifier trained on other targets could not beat the Majority baseline on this test set.

6.2 Task B Participating Stance Systems

Nine teams competed in Task B. Table 4 shows each team's performance. Teams are sorted in terms of their performance according to the official metric. The best results obtained by a participating system was an F_{avg} of 56.28 by *pkudblab*. They used a rule-based annotation of the domain corpus to train a deep convolutional neural network to differentiate 'favour' from 'against' instances. At test time, they combined their network's output with rules to produce predictions that include the 'neither' class.

In general, results for Task B are lower than those for Task A as one would expect, as we remove the benefit of direct supervision. However, they are perhaps not as low as we might have expected, with the best result of 56.28 actually beating the best result for the supervised 'Climate Change' task (54.86).

Team	**Opinion Towards** Target	Other	**All**
Baselines			
Majority class	35.20	25.52	29.72
SVM-ngrams-comb	31.39	20.13	28.43
Participating Teams			
pkudblab	**67.19**	25.77	**56.28**
LitisMind	51.60	**29.50**	44.66
INF-UFRGS	50.04	22.66	42.32
UWB	50.62	25.02	42.02
ECNU	40.66	19.14	34.08
USFD	38.87	22.80	32.70
Thomson Reuters	38.06	22.60	32.39
ltl.uni-due	34.16	4.69	26.14
NEUSA	28.86	18.35	25.73

Table 5: Results for Task B (the official competition metric F_{avg}) on different subsets of the test data. The highest score in each column is shown in bold.

Table 5 shows results for Task B on subsets of the test set where opinion is expressed towards the target of interest and towards some other entity. Observe that here too results are markedly lower when stance is to be inferred from a tweet expressing opinion about some other entity (and not the target of interest).

6.3 Discussion

Some teams did very well detecting tweets in favor of Trump (*ltl.uni-due*), with most of the others performing best on tweets against Trump. This makes sense, as 'against' tweets made up the bulk of the Trump dataset. The top team, *pkudblab*, was the only one to successfully balance these two goals, achieving the best F_{favor} score and the second-best $F_{against}$ score.

The Task B teams varied wildly in terms of approaches to this problem. The top three teams all took the approach of producing noisy labels, with *pkudblab* using keyword rules, *LitisMind* using hashtag rules on external data, and *INF-UFRGS* using a combination of rules and third-party sentiment classifiers. However, we were pleased to see other teams attempting to generalize the supervised data from Task A in interesting ways, either using rules or multi-stage classifiers to bridge the target gap. We are optimistic that there is much interesting follow-up work yet to come on this task.

Further details on the submissions can be found in the system description papers published in the SemEval-2016 proceedings, including papers by Elfardy and Diab (2016) for *CU-GWU*, Dias and Becker (2016) for *INF-URGS*, Patra et al. (2016) for *JU_NLP*, Wojatzki and Zesch (2016) for *ltl.uni-due*, Zarrella and Marsh (2016) for *MITRE*, Misra et al. (2016) for *nldsucsc*, Wei et al. (2016) for *pkudblab*, Tutek et al. (2016) for *TakeLab*, Yuki et al. (2016) for *Tohoku*, and Augenstein et al. (2016) for *USFD*.

7 Related Work

Past work on stance detection includes that by Somasundaran and Wiebe (2010), Anand et al. (2011), Faulkner (2014), Rajadesingan and Liu (2014), Djemili et al. (2014), Boltuzic and Šnajder (2014), Conrad et al. (2012), Sridhar et al. (2014), Rajadesingan and Liu (2014), and Sobhani et al. (2015). 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; Mohammad et al., 2013; Rosenthal et al., 2015). See Pontiki et al. (2014), Pontiki et al. (2015), and Kiritchenko et al. (2014a) for tasks and systems on aspect based sentiment analysis, where the goal is to determine sentiment towards aspects of a product such as speed of processor and screen resolution of a cell phone.

8 Conclusions and Future Work

We described a new shared task on detecting stance towards pre-chosen targets of interest from tweets. We formulated two tasks: a traditional supervised task where labeled training data for the test data targets is made available (Task A) and a more challenging formulation where no labeled data pertaining to the test data targets is available (Task B). We received 19 submissions for Task A and 9 for Task B, with systems utilizing a wide array of features and resources. Stance detection, especially as formulated for Task B, is still in its infancy, and we hope that the dataset made available as part of this task will foster further research not only on stance detection as proposed here, but also for related tasks such as exploring the different ways in which stance is conveyed, and how the distribution of stance towards a target changes over time.

Acknowledgment

Annotations of the Stance Dataset were funded by the National Research Council of Canada.

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 2nd workshop on computational approaches to subjectivity and sentiment analysis*, pages 1–9.

Isabelle Augenstein, Andreas Vlachos, and Kalina Bontcheva. 2016. USFD at SemEval-2016 Task 6: Any-Target Stance Detection on Twitter with Autoencoders. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, June.

Filip Boltuzic and Jan Šnajder. 2014. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58.

Alexander Conrad, Janyce Wiebe, and Rebecca Hwa. 2012. Recognizing arguing subjectivity and argument tags. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 80–88.

Marcelo Dias and Karin Becker. 2016. INF-UFRGS-OPINION-MINING at SemEval-2016 Task 6: Automatic Generation of a Training Corpus for Unsupervised Identification of Stance in Tweets. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, June.

Sarah Djemili, Julien Longhi, Claudia Marinica, Dimitris Kotzinos, and Georges-Elia Sarfati. 2014. What does Twitter have to say about ideology? In *Proceedings of the Natural Language Processing for Computer-Mediated Communication/Social Media-Pre-conference workshop at Konvens*.

Heba Elfardy and Mona Diab. 2016. CU-GWU at SemEval-2016 Task 6: Perspective at SemEval-2016 Task 6: Ideological Stance Detection in Informal Text. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, June.

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 Twenty-Seventh International Flairs Conference*.

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

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*, SemEval '14, 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.

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 US.

Amita Misra, Brian Ecker, Theodore Handleman, Nicolas Hahn, and Marilyn Walker. 2016. nldsucsc at SemEval-2016 Task 6: A Semi-Supervised Approach to Detecting Stance in Tweets. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, June.

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 NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34.

Saif 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*, SemEval '13, 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 10th edition of the the Language Resources and Evaluation Conference (LREC)*, Portorož, Slovenia.

Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2016b. Stance and sentiment in tweets. *Special Section of the ACM Transactions on Internet Technology on Argumentation in Social Media*, Submitted.

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

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

Braja Gopal Patra, Dipankar Das, and Sivaji Bandyopadhyay. 2016. JU_NLP at SemEval-2016 Task 6: Detecting Stance in Tweets using Support Vector Machines. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, June.

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*, SemEval '14, 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*, SemEval '15, 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.

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 9th 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. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124.

Dhanya Sridhar, Lise Getoor, and Marilyn Walker. 2014. Collective stance classification of posts in online debate forums. *Proceedings of the Association for Computational Linguistics*, page 109.

Martin Tutek, Ivan Sekulić, Paula Gombar, Ivan Paljak, Filip Čulinović, Filip Boltužić, Mladen Karan, Domagoj Alagić, and Jan Šnajder. 2016. TakeLab at SemEval-2016 Task 6: Stance Classification in Tweets Using a Genetic Algorithm Based Ensemble. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, June.

Wan Wei, Xiao Zhang, Xuqin Liu, Wei Chen, and Tengjiao Wang. 2016. pkudblab at SemEval-2016 Task 6: A Specific Convolutional Neural Network System for Effective Stance Detection. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, June.

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*, SemEval '13, Atlanta, Georgia, USA, June.

Michael Wojatzki and Torsten Zesch. 2016. ltl.uni-due at SemEval-2016 Task 6: Stance Detection in Social Media Using Stacked Classifiers. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, June.

Igarashi Yuki, Komatsu Hiroya, Kobayashi Sosuke, Okazaki Naoaki, and Inui Kentaro. 2016. Tohoku at SemEval-2016 Task 6: Feature-based Model versus Convolutional Neural Network for Stance Detection. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, June.

Guido Zarrella and Amy Marsh. 2016. MITRE at SemEval-2016 Task 6: Transfer Learning for Stance Detection. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, June.