A Dataset for Detecting Stance in Tweets

Saif M. Mohammad[1], **Svetlana Kiritchenko**[1], **Parinaz Sobhani**[2], **Xiaodan Zhu**[1], **Colin Cherry**[1]
[1]National Research Council Canada, [2]University of Ottawa
saif.mohammad@nrc-cnrc.gc.ca, svetlana.kiritchenko@nrc-cnrc.gc.ca, psobh090@uottawa.ca,
xiaodan.zhu@nrc-cnrc.gc.ca, colin.cherry@nrc-cnrc.gc.ca

Abstract

We can often detect from a person's utterances whether he/she is in favor of or against a given target entity (a product, topic, another person, etc.). 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 of interest—their stance. The targets of interest may or may not be referred to in the tweets, and they may or may not be the target of opinion in the tweets. The data pertains to six targets of interest commonly known and debated in the United States. Apart from stance, the tweets are also annotated for whether the target of interest is the target of opinion in the tweet. The annotations were performed by crowdsourcing. Several techniques were employed to encourage high-quality annotations (for example, providing clear and simple instructions) and to identify and discard poor annotations (for example, using a small set of check questions annotated by the authors). This Stance Dataset, which was subsequently also annotated for sentiment, can be used to better understand the relationship between stance, sentiment, entity relationships, and textual inference.

Keywords: Stance, argumentation, tweets, sentiment, opinion, target of opinion

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, 2013b; Sridhar et al., 2014).

The task we explore is detecting stance from tweets, and it is formulated as follows: given a tweet and a target entity (person, organization, movement, policy, etc.), can automatic natural language systems 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
> Tweet: *A foetus has rights too! Make your voice heard.*

Humans can deduce from the tweet that the tweeter is likely against the target (the tweeter's stance is against the legalization of abortion). Our goal is to create labeled training and test data that can be used in the developments of automatic systems for detecting stance.

In this paper, we describe how we created a dataset of 4,870 tweet–target pairs that are manually annotated for stance. We will refer to this data as the *Stance Dataset*. The dataset has instances corresponding to six pre-chosen targets of interest: 'Atheism', 'Climate Change is a Real Concern', 'Feminist Movement', 'Hillary Clinton', 'Legalization of Abortion', and 'Donald Trump'. The annotations were performed by crowdsourcing. Several techniques were employed to encourage high-quality annotations and to identify and discard poor annotations.

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 of neutral stance is shown below:

> Target: Hillary Clinton
> Tweet: *Hillary Clinton has some strengths and some weaknesses, I could vote either way come election day.*

Thus, in our work we obtain manual annotations for 'favor', 'against', 'neutral', and 'no stance', but later combine 'neutral' and 'no stance' into one category 'neither' (neither favor nor against) since less than 0.1% of the data received the 'neutral' label.

To successfully detect stance, automatic systems often have to rely on world knowledge that may not be explicitly stated in the focus text. For example, systems benefit from knowing that if one is actively supporting foetus rights, then he or she is likely against the right to abortion. This world knowledge may be acquired from large text corpora. Thus for each target, we also acquire a corpus of unlabeled tweets that include hashtags related to the target. We will refer to this set of tweets as the domain corpus for the target. Automatic systems can gather information from the domain corpus to help with the detection of stance—for example, by identifying how entities are related.

Stance detection is related to sentiment analysis, but the two have significant differences. In sentiment analysis, systems

determine whether a piece of text is positive, negative, or neutral. However, in stance detection, systems are to determine favorability towards a given target of interest—and the target may not be explicitly mentioned in the text. For example, consider the target–text pair below:

> Target: Donald Trump
> Text: *Jeb Bush is the only sane candidate in this republican lineup.*

The target of opinion in the tweet is Jeb Bush, but the given stance target is Donald Trump. The tweet expresses positive opinion towards Jeb Bush, from which we can infer that the tweeter is likely to be unfavorable towards Donald Trump. Note that it is possible that one can be in favor of Jeb Bush and yet also be in favor of Donald Trump. However, the goal in stance detection, is to determine which is more probable: that the author is in favor of, against, or neutral towards the target. In this case, most annotators will agree that the tweeter is likely against Donald Trump. To aid further analysis, the tweets in the Stance Dataset are also annotated for whether target of interest is the target of opinion in the tweet.

Partitions of the Stance Dataset were used to create training and test sets for the SemEval-2016 Task 6: Detecting Stance from Tweets (Mohammad et al., 2016a).[1] Mohammad et al. (2016b) subsequently annotated the Stance Dataset for sentiment and quantitatively explored the relationship between stance and sentiment.

The rest of the paper is structured as follows. In Section 2, we describe how we created the Stance Dataset. Section 3 presents a detailed analysis of the stance annotations. Section 4 presents an online interactive visualization of the Stance Dataset. Section 5 discusses how the dataset can be (and is being) used by the research community. Finally we present concluding remarks in Section 6. All of the data created as part of this project (the Stance Dataset, the domain corpus, the annotation questionnaire, etc.) as well as an interactive visualization to explore the data are made freely available.[2]

2. Creating the Dataset for Stance in Tweets

In order to create a suitable dataset of tweet–target pairs annotated for stance, we first identified useful properties for such a dataset (Section 2.1), then selected tweet–target pairs in a manner that is consistent with those properties (Section 2.2), and finally annotated the tweet–target pairs using a carefully developed set of instructions and questionnaire (Section 2.3).

2.1. Properties of a Good Stance-Labeled Dataset

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

1: *The tweet and target are commonly understood by a wide number of people in the United States.*
This is important because the tweet-target pairs will later be annotated by English speakers from the United States.

2: *There must be a significant amount of data for each of the three classes: favor, against, neither.*
Often, the proportion of tweets in favor of a target may not be similarly numerous as those against it. However, we did not want scenarios where there are no tweets in favor of a target or no tweets against it. Also, the total number of tweet–target pairs from which the stance cannot be inferred ('neither' instances) can be very large. However, creating a dataset where 99% of the tweets are from this category makes the dataset less interesting and less useful. So we down-sample the number of 'neither' instances.

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.*
We wanted to include the relatively harder cases for stance detection where the target is referred to in indirect ways such as through pronouns, epithets, honorifics, and relationships.

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 stance target.*
As mentioned earlier with the Donald Trump example, sometimes stance towards a target can be inferred even if that target is not the target of opinion in the text. Including such instances makes the task more challenging. Downstream applications often require stance towards particular pre-chosen targets, and having data where the target of opinion is different from the target of stance helps test how well stance detection systems can cope with such instances.

These properties influenced various choices in how our dataset was created.

2.2. Selecting the Tweet–Target Pairs for Stance Annotation

There are two broad ways in which the tweet–target pairs could be obtained:

- *Random search*: Poll the Twitter API for a random selection of tweets for some period of time. Manually identify targets towards whom stance can be determined.
- *Targeted search*: First identify a list of potential targets. Poll the Twitter API for tweets relevant to these targets.

As mentioned above in reference to Property 1, one of the challenges with creating good data for stance detection is that human annotators can find it difficult to determine stance if they do not understand the domain or the relationships between relevant entities. Thus, we chose the targeted search option as this allowed us to focus on only those targets that are widely known. Additionally, this approach allows creation of many instances for each of the targets. This

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

Target	Example Favor Hashtag	Example Against Hashtag	Example Stance-Ambiguous Hashtag
Atheism	*#NoMoreReligions*	*#Godswill*	*#atheism*
Climate Change Concern	-	*#globalwarminghoax*	*#climatechange*
Donald Trump	#Trump2016	-	#WakeUpAmerica
Feminist	*#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.

is significant because stance detection systems often create separate models for each target using labeled training data. The authors of this paper selected as targets a small subset of entities routinely discussed on Twitter at the time of data collection: 'Atheism', 'Climate Change is a Real Concern', 'Feminist Movement', 'Hillary Clinton', 'Legalization of Abortion', and 'Donald Trump'.

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*).[3] We will refer to favor and against hashtags jointly as *stance-indicative (SI) hashtags*. Table 1 lists some of the hashtags used for each of the targets. (We were not able to find a hashtag that is predominantly used to show favor towards 'Climate change is a real concern', however, the stance-ambiguous hashtags were the source of a large number of tweets eventually labeled 'favor' through human annotation.) 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. For human annotation of stance, for each target, we sample an equal number of tweets pertaining to the favor hashtags, the against hashtags, and the stance-ambiguous hashtags. This encourages (but does not guarantee) a more equitable number of tweets pertaining to the stance categories (Property 2).

Note that the tweets that have a favor hashtag may oppose the target as well. Further, once the favor hashtag is removed, the tweet may not have enough information to suggest that the tweeter is favorable towards the target. The same is true for tweets obtained with the against hashtags. Thus, our procedure for obtaining tweets results in tweets with various stance class distributions.

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
> Tweet: *Benghazi questions need to be answered #Jeb2016 #HillNo*

Removal of *#HillNo* leaves no mention of Hillary Clinton, but yet there is sufficient evidence (through references to Benghazi and #Jeb2016) that the tweeter is 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.

2.3. Stance Annotation

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. Thus after a few rounds of internal development and pilot annotations, we presented the questionnaire shown below to the annotators. Apart from a question on stance, we also asked a second question pertaining to whether the target of opinion in the tweet is the same as the target, some other entity, or neither.

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

   This could be because of any of reasons shown below:
   - *the tweet is explicitly in support for the target*
   - *the tweet is in support of something/someone aligned with the target, from which we can infer that the tweeter supports the target*
   - *the tweet is against something/someone other than the target, from which we can infer that the tweeter supports the target*
   - *the tweet is NOT in support of or against anything, but it has some information, from which we can infer that the tweeter supports the target*
   - *we cannot infer the tweeter's stance toward the target, but the tweet is echoing somebody else's favorable stance towards the target (this could be a news story, quote, retweet, etc)*

2. We can infer from the tweet that the tweeter is against the target

   This could be because of any of the following:

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

- *the tweet is explicitly against the target*
- *the tweet is against someone/something aligned with the target entity, from which we can infer that the tweeter is against the target*
- *the tweet is in support of someone/something other than the target, from which we can infer that the tweeter is against the target*
- *the tweet is NOT in support of or against anything, but it has some information, from which we can infer that the tweeter is against the target*
- *we cannot infer the tweeter's stance toward the target, but the tweet is echoing somebody else's negative stance towards the target entity (this could be a news story, quote, retweet, etc)*

3. We can infer from the tweet that the tweeter has a neutral stance towards the target

   The tweet must provide some information that suggests that the tweeter is neutral towards the target – the tweet being neither favorable nor against the target is not sufficient reason for choosing this option. One reason for choosing this option is that the tweeter supports the target entity to some extent, but is also against it to some extent.

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 about anything

---

For each of the six 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.[4] 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. Table 2 shows the number of tweets available for each target in the domain corpus.

3. Analysis of the Annotations

We compared responses to the stance question (Q1) from each crowd annotator with gold labels in a small dataset of internally annotated instances. If a crowd annotator's responses did not match the gold labels for at least 70% of the instances, then all of their responses were discarded. The inter-annotator agreement on the remaining stance responses was about 73.11%. These 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).

Since we wanted to create a dataset for training and testing of automatic stance detection systems, we use only those instances for which inter-annotator agreement was greater than 60%. That is, we include only those instances for which the majority stance label is chosen by at least 60% of the annotators.[5] This resulted in a dataset of 4,870 instances labeled for stance. Table 3 shows the number of tweets per target. The break down of these tweets into training and test sets, as well as the distributions for favor, against, and neither, are shown in Section 4.1, where we discuss how the dataset was used in a SemEval-2016 shared task.

Table 4 shows the distribution of responses to Question 2 (whether opinion is expressed directly about our target, about somebody/someone other than the target, or no opinion is being expressed). Observe that the percentage of 'opinion towards other' varies across different targets from 30% to 50%. Inter-annotator agreement for responses to Question 2 was 68.90%. Table 5 shows the distribution of instances by target of opinion, for each of the stance labels. Observe that 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 28%).

After some initial annotations, we examined instances pertaining to the targets 'Hillary Clinton' and 'Legalization of Abortion' to identify tweets that do not mention the target explicitly, but yet 'favor' or 'against' stance can be inferred. We wanted to determine whether our data has instances where stance towards the target can be inferred even though the target is not explicitly mentioned. Some examples are shown below.

Target	# Tweets
Atheism	935,181
Climate Change Concern	208,880
Donald Trump	78,156
Feminist Movement	144,166
Hillary Clinton	238,193
Legalization of Abortion	113,193
Total	1,717,769

Table 2: Number of tweets in the domain corpus.

Target	**# instances**
Atheism	733
Climate Change Concern	564
Donald Trump	707
Feminist Movement	949
Hillary Clinton	984
Legalization of Abortion	933
Total	4870

Table 3: Number of instances labeled for stance.

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

[5] This is a somewhat arbitrary threshold, but it seemed appropriate in terms of balancing confidence in the majority annotation and having to discard too many instances. Annotations for about 25% of the instances do not satisfy this criterion.

	Opinion towards		
Target	Target	Other	No one
Atheism	49.25	46.38	4.37
Climate Change Concern	60.81	30.50	8.69
Donald Trump	45.83	50.35	3.82
Feminist Movement	68.28	27.40	4.32
Hillary Clinton	60.32	35.10	4.58
Legalization of Abortion	63.67	30.97	5.36
Total	58.80	36.19	5.01

Table 4: Distribution of target of opinion.

	Opinion towards		
Stance	Target	Other	No one
For	94.69	4.73	0.58
Against	71.01	28.32	0.66
Neither	0.95	81.45	17.60

Table 5: Distribution (in %) of target of opinion by stance.

For target 'Hillary Clinton':
Tweet: *I think I am going to vote for Monica Lewinsky's Ex-boyfriends Wife*
Tweet: *Let's hope the VOTERS remember! #HillNo*
Tweet: *How can she live with herself? #Benghazi*

For target 'Legalization of Abortion':
Tweet: *Why dehumanize the pregnant person? They're more than walking incubators, and have rights!*
Tweet: *the woman has a voice the doctor has a voice. Who speaks for the baby? I'm just askin.*
Tweet: *Today I am grateful to have the right to control my body without govt influence. #abvote*

In all, about 30% of the 'Hillary Clinton' instances and about 65% of the 'Legalization of Abortion' instances were found to be of this kind—that is, 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). This marked proportion of instances that do not explicitly refer to the target of interest makes the Stance Dataset a particularly challenging, but realistic, test set for stance classification.

4. An Interactive Visualization of Stance and Sentiment

An interactive visualization of the Stance Dataset that shows various statistics about the data is made available online.[6] Figure 1 is a screenshot of the home screen. Note that the visualization also shows sentiment and target of opinion annotations (in addition to stance). On the top left is a bar graph showing the number of instances pertaining to each of the targets in the dataset. The visualization component below it, known as a treemap, shows tiles corresponding to each target–stance combination. The size (area) of a tile is proportional to the number of instances corresponding to that target–stance combination. This component shows that for most of the targets, the Stance Dataset has more data for 'against' than 'favor' and 'neither'. The three stacked bars on the top right show the proportion of instances pertaining to the classes of stance, opinion target, and polarity, respectively. Observe that they convey to the viewer that a majority of the instances are labeled as 'against' the targets of interest, expressing opinion towards the target of interest, and having negative polarity.

The 'X by Y Matrices' component of the visualization shows three matrices pertaining to: stance classes and opinion towards classes, stance classes and polarity classes, and opinion towards classes and polarity classes. The cells in each of these matrices show the percentage of instances with labels corresponding to that cell (the percentages across each of the rows sums up to 100%.) For example, observe in the left-most matrix that favorable stance is usually expressed by providing opinion directly about the target (94.23%), but that percentage is markedly smaller for instances that are labeled 'against the target' (72.75%). The visualization component at the bottom shows all of the tweets, targets, and manual annotations.

Clicking on visualization elements filters the data. For example, clicking on 'Feminism' and 'Favor' will show information pertaining to 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.

5. Applications of the Stance Dataset

The Stance Dataset is already being used by the research community for several purposes. Here we describe a few of the current and possible future applications of the data.

5.1. SemEval-2016 Task 6: Detecting Stance in Tweets

The Stance Dataset was used as the official training and test data in the SemEval-2016 shared task on Detecting Stance in Tweets (Task 6) (Mohammad et al., 2016a).[7] Submissions were solicited in two formulations (Task A and Task B). 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. Table 6 shows the distribution of stance labels in the training and test sets.

Task A received submissions from 19 teams, wherein the highest classification F-score obtained was 67.8. Task B, which is particularly challenging due to lack of training data, received submissions from 9 teams wherein the highest F-score obtained was 56.3. The best performing systems used standard text classification features such as those drawn from ngrams, word vectors, and sentiment lexicons such as the NRC Emotion Lexicon (Mohammad and Turney, 2013). Some teams drew additional gains from noisy stance-labeled data created using distant supervision techniques. Mohammad et al. (2016b) proposed a method for stance detection using various surface-form features, word embeddings, and distant supervision that obtained even better F-scores (close to 70.0).

[6] http://www.saifmohammad.com/WebPages/StanceDataset.htm

[7] http://alt.qcri.org/semeval2016/task6/

Target	# total	# train	% of instances in Train favor	against	neither	# test	% of instances in 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 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 6: Distribution of instances in the stance datasets used in SemEval-2016 Task 6 Task A and Task B.

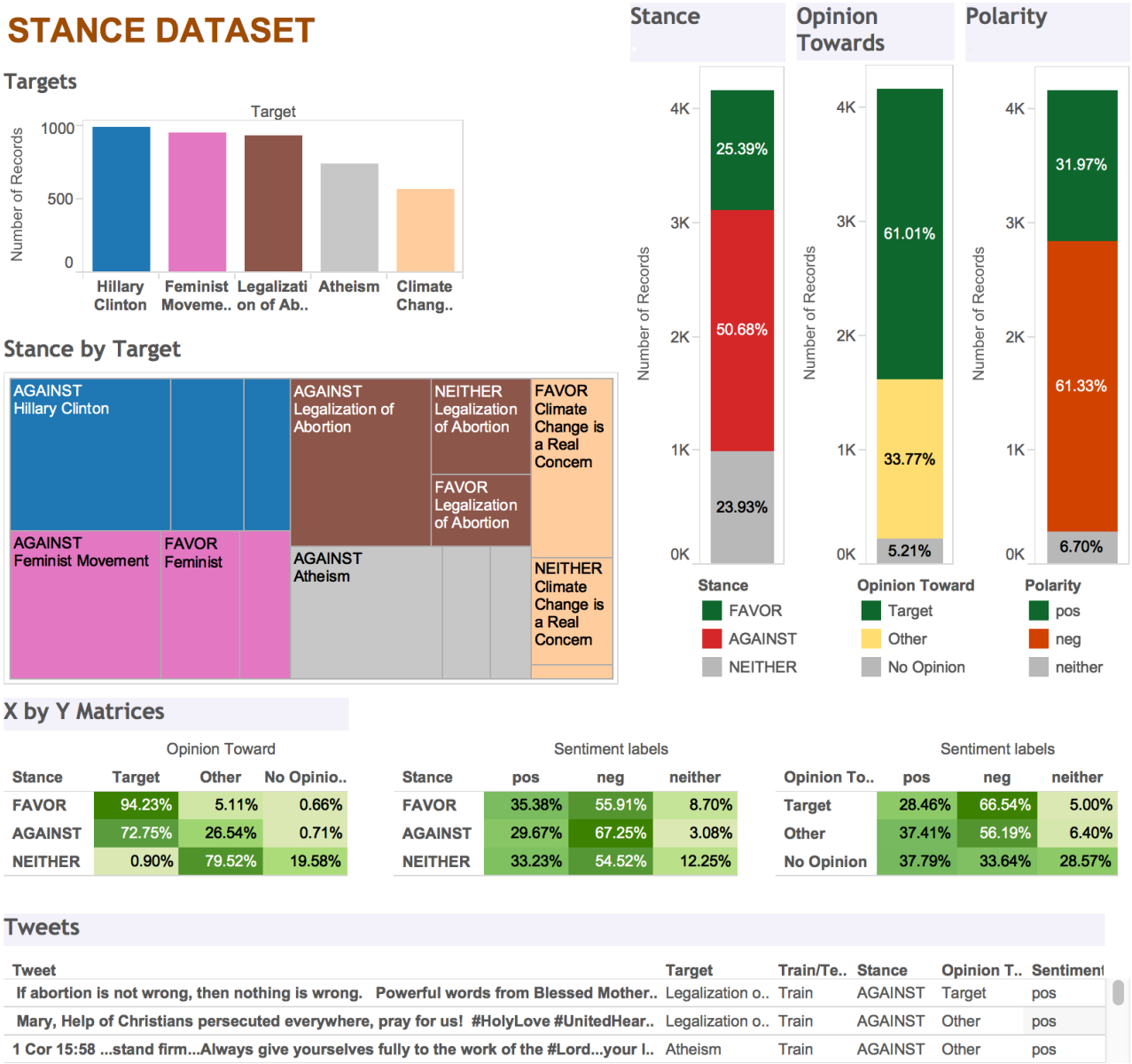

Figure 1: Screenshot of an Interactive Visualization of the Stance Dataset. On the top left is a bar graph showing the number of instances pertaining to each of the targets in the dataset. The visualization component below it, known as a treemap, shows tiles corresponding to each target–stance combination. The size (area) of a tile is proportional to the number of instances corresponding to that target–stance combination. The 'X by Y Matrices' component of the visualization shows three matrices pertaining to: stance classes and opinion towards classes, stance classes and polarity classes, and opinion towards classes and polarity classes. The table at the bottom shows all tweets pertaining to the current selection.

5.2. Understanding the Relationship and Interaction between Stance and Sentiment

Mohammad et al. (2016b) annotated the tweets in the Stance Dataset for whether they convey positive, negative, or neutral sentiment. They conducted an experiment using the manual labels to determine the extent to which stance can be determined simply from sentiment. They also built a common text classification framework that relies on a variety of features, including those drawn from sentiment lexicons, to determine both stance and sentiment. The results show that while sentiment features are useful for stance detection, they alone are not sufficient. 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. They also 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.

Finally, they conduct several experiments exploring the use of distant supervision for stance detection. They determine the extent to which removal of stance-indicative hashtags from tweets still leaves sufficient information in the tweet to convey that the stance of the tweeter is the same that indicated by removed hashtag. They also create a subset of the domain corpus that is pseudo-labeled for stance and show that stance classification systems can benefit from using it (either as additional training data or by extracting features from the data).

5.3. Other Applications

The Stance Dataset can be extended to more targets and domains using the same general methodology used to create it. Thus, the dataset and its extensions can be used in a number of applications such as detecting stance towards politicians, products, government policies, social issues, and so on. One approach to detecting stance is identifying relationships between entities. For example, knowing that entity X is an adversary of entity Y can be useful in detecting stance towards Y in tweets that mention X (and not Y). Thus the stance dataset and the associated classification task can be used for developing and evaluating automatic approaches for relationship extraction. Since the Stance Dataset is also annotated for target of opinion, it can be used to better understand how stance can be detected from tweets that do not explicitly mention the target of interest. Stance detection can be thought of as a textual inference or textual entailment task, where the goal is to determine whether the favorability of the target is entailed by the tweet. Thus the dataset can be used for developing textual inference engines and open domain reasoning.

6. 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), Hasan and Ng (2013a), Djemili et al. (2014), Sridhar et al. (2014), and Sobhani et al. (2015). In one of the few works on stance detection in tweets, Rajadesingan and Liu (2014) determine 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. Djemili et al. (2014) use a set of rules based on the syntax and discourse structure of the tweet to identify tweets that contain ideological stance. However, none of these works attempts to determine stance from a single tweet.

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). 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; Kiritchenko et al., 2014; Pontiki et al., 2014).

7. Summary

We presented a new dataset of 4,870 tweet–target pairs annotated for stance of the tweeter towards the target. This dataset, which we refer to as the *Stance Dataset*, has instances corresponding to six pre-chosen targets of interest: 'Atheism', 'Climate Change is a Real Concern', 'Feminist Movement', 'Hillary Clinton', 'Legalization of Abortion', and 'Donald Trump'. The annotations were performed by crowdsourcing. Several techniques were employed to encourage high-quality annotations and to identify and discard poor annotations. We analyzed the dataset to show that it has several interesting properties. For example, a marked number of tweets do not explicitly mention the target, and in many tweets the target of opinion is different from the given target of interest. Mohammad et al. (2016b) subsequently annotated the Stance Dataset for sentiment and quantitatively explored the relationship between stance and sentiment.

The Stance Dataset can be extended to more targets and domains using the same general methodology used to create it. Thus, the dataset and its extensions can be used in a number of applications such as tracking sentiment towards politicians, products, and issues. Partitions of the Stance Dataset were used as the official test and training sets in the SemEval-2016 Task 6: Detecting Stance from Tweets. The shared task received more than 25 submissions across two variants of stance detection tasks. Mohammad et al. (2016b) proposed a method for stance detection using various surface-form features, word embeddings, and distant supervision that obtained even better F-scores than the best participating team in SemEval-2016 Task 6. All of the data created as part of this project (the Stance Dataset, the domain corpus, the annotation questionnaire, etc.) as well as an interactive visualization to explore the data are made freely available.[8] We hope this will encourage more work that brings together the fields of sentiment analysis, textual inference, and relationship extraction.

[8] http://www.saifmohammad.com/WebPages/StanceDataset.htm

8. Bibliographic References

Anand, P., Walker, M., Abbott, R., Tree, J. E. F., Bowmani, R., and Minor, M. (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.

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

Conrad, A., Wiebe, J., and Hwa, R. (2012). Recognizing arguing subjectivity and argument tags. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 80–88.

Djemili, S., Longhi, J., Marinica, C., Kotzinos, D., and Sarfati, G.-E. (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*.

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

Hasan, K. S. and Ng, V. (2013a). Extra-linguistic constraints on stance recognition in ideological debates. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 816–821.

Hasan, K. S. and Ng, V. (2013b). Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1348–1356.

Kiritchenko, S., Zhu, X., Cherry, C., and Mohammad, S. M. (2014). NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '14, Dublin, Ireland, August.

Liu, B. and Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In Charu C. Aggarwal et al., editors, *Mining Text Data*, pages 415–463. Springer US.

Mohammad, S. M. and Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Mohammad, S., Kiritchenko, S., and Zhu, X. (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.

Mohammad, S. M., Kiritchenko, S., Sobhani, P., Zhu, X., and Cherry, C. (2016a). SemEval-2016 Task 6: Detecting stance in tweets. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, June.

Mohammad, S. M., Sobhani, P., and Kiritchenko, S. (2016b). Stance and sentiment in tweets. *Special Section of the ACM Transactions on Internet Technology on Argumentation in Social Media*, Submitted.

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

Murakami, A. and Raymond, R. (2010). Support or oppose? Classifying positions in online debates from reply activities and opinion expressions. In *Proceedings of Coling 2010*, pages 869–875, Beijing, China.

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

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., and Manandhar, S. (2014). SemEval-2014 Task 4: Aspect based sentiment analysis. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '14, Dublin, Ireland, August.

Pontiki, M., Galanis, D., Papageogiou, H., Manandhar, S., and Androutsopoulos, I. (2015). SemEval-2015 Task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, Colorado*.

Rajadesingan, A. and Liu, H. (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.

Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S. M., Ritter, A., and Stoyanov, V. (2015). Semeval-2015 Task 10: Sentiment analysis in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluations*.

Sobhani, P., Inkpen, D., and Matwin, S. (2015). From argumentation mining to stance classification. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 67–77, Denver, Colorado, USA.

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

Somasundaran, S. and Wiebe, J. (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.

Sridhar, D., Getoor, L., and Walker, M. (2014). Collective stance classification of posts in online debate forums. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, page 109.

Thomas, M., Pang, B., and Lee, L. (2006). Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 327–335.

Walker, M. A., Anand, P., Abbott, R., and Grant, R. (2012). Stance classification using dialogic properties of persuasion. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 592–596.

Wilson, T., Kozareva, Z., Nakov, P., Rosenthal, S., Stoyanov, V., and Ritter, A. (2013). SemEval-2013 Task 2: Sentiment analysis in Twitter. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '13, Atlanta, Georgia, USA, June.