Yuxuan Xu. Twitter Stance Detection with Textual, Sentiment, and Target-specific Models. A Master's Paper for the M.S. in I.S degree. April, 2020. 42 pages. Advisor: Jaime Arguello

Today more and more users express their opinions and stances on social media platforms such as Twitter. In this paper, I proposed different approaches to automatically detect the stance of a single tweet. I investigated whether including additional sentiment polarity information and the target information would be beneficial for the stance detection task. Moreover, I also researched whether target-specific features could be generalized to other datasets with different targets for the stance detection task.

Headings:

Natural language processing

Text-based prediction

Stance detection

TWITTER STANCE DETECTION WITH TEXTUAL, SENTIMENT, AND TARGET-SPECIFIC MODELS

by
Yuxuan Xu

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

April 2020

Approved by

_____

Jaime Arguello

Table of Contents

Introduction

With the booming of technology in this information era, the capacity of data and information are growing exponentially on a daily basis. People have numerous sources for acquiring information through various formats: cell phones, traditional media such as TV and magazines, podcasts, and entertainment like movies and TV series. In addition, people feel that it is necessary to let their voices be heard and share their opinion towards almost every aspect of their life on social media such as Twitter, Facebook and online forums like Reddit.

As defined in the Merriam Webster dictionary, a stance is "an intellectual or emotional attitude" (*Definition of STANCE*, n.d.). It is an attitude a person holds towards a certain topic or target. According to Mohammad et al. (2017, p. 1), stance detection is "the task of automatically determining from the text whether the author of the text is in favor of, against, or neutral towards a proposition or target". This target could be a political figure, a theory, a movement, or a product, etc. For example, one person could take a stance for the existence of God or against the existence of God. Traditionally, people tend to have a debate or speech for such issues and therefore the format of such debate would be a long discussion. This could happen during an actual face-to-face debate or on an online discussion forum. People nowadays would also leave their opinion, whether they are in favor of, against, or neutral towards a particular target on online platforms. Specifically,

on Twitter, a social media platform, users could post a short tweet due to the character limit (*Counting characters*, n.d.). Rather than a tedious statement, Twitter users have the opportunity to state their stance and argue for it through a relatively short post.

For example, on the topic of "Legalization of Abortion", one Twitter user wrote, "We live in a world where people care more about religious rights than consent & reproductive rights....". The user explicitly supports the legalization of abortion while expressing negative sentiment. The user's negative sentiment isn't directed toward the target issue while expressing an unfriendly tone. The sentiment involved in such a statement has been a major field of study for natural language processing and text mining, however, the detection of the stance its writer took hasn't received much attention.

The paper first conducted a literature review on the development of stance detection, its relationship with sentiment analysis, and previous work on the topic of stance detection. Then this paper introduced a method to use tweets to predict its stance toward a particular target, whether sentiment expressed is beneficial for stance detection, and determine if the results vary among different topics.

2 Literature Review

This section presents the review of prior studies and literature on this research. It starts with an introduction of sentiment analysis and its association with stance detection, then moves into a more detailed review of stance detection based on the application areas. Finally, despite various application fields, a summary of models used across different domains will be offered, followed by a list of research questions for this study.

2.1 Sentiment analysis and its relationship with Stance detection

Stance detection is associated with, but different from, sentiment analysis. In the field of NLP (natural language processing), sentiment analysis came along with the development of the Internet and Web 2.0 (Du et al., n.d.). And with the abundance of online text, according to Mohammad et al., sentiment analysis tasks are 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" (2017, p. 2). The phrase "sentiment analysis" is often exchangeable with the term "opinion mining" in the academic field (Pang & Lee, 2008; Patodkar & I.R, 2016). It mainly focuses on comprehending what users think about a certain topic or text (Cabrio & Villata, 2018, p. 5428). For instance, document-level opinion mining could help with determining the polarity expressed in a customer review (Liu, 2012; Pang & Lee, 2008).

A significant difference between the task of stance detection and that of sentiment analysis is that stance detection is dependent on both the subjective expression found in the text and the associated target of stance detection tasks might not be explicitly mentioned in the text (Du et al., n.d.). Mohammad et al. also points out that in stance detection, systems are to determine favorability towards a given target of interest, and this target of interest might not be the target of opinion in the text (2017, p. 2).

Unlike sentiment analysis, which has gained researchers' attention for a relatively long time, stance detection is one of the researchers' new interests in recent years. Depending on the categories of text, stance detection could be applied to various domains such as education, legal documents, political debates and speeches, and web-based content such as microblogs and online product reviews (Cabrio & Villata, 2018).

2.2 Stance detection categorization based on application areas

In the following section, previous literature will be presented based on the content type that they focused on. First, it introduced previous work of stance detection in the field of online discussions, debates, and speeches. Then it moved on to prior studies for two specific challenges, Fake New Challenge and Rumor Detection Challenge. Finally, previous research on stance detection of tweets will be introduced, which is the more relevant area to this study.

2.2.1 Stance detection for discussion, debates, and speeches

Political debates and speeches are the first of several areas that scholars studied since they are usually longer in context and full of arguments and premises for various purposes. Due to their length and context, according to Lippi and Torroni, application in the political domain could extend from simple stance detection to the final aim of detecting fallacies, persuasiveness degree, and coherence in the candidate's argumentation (2016). Rory et al. studied the presence and polarity of ethotic arguments from UK parliamentary debates (2016). Naderi and Hirst (2016) proposed a corpus of speeches from the Canadian Parliament and examined the statements with respect to the position of the speaker towards the discussed topic (as cited in Cabrio & Villata, 2018).

Rather than an online debate forum, Levow et al. exploited the ATAROS corpus, a corpus of task-oriented spontaneous speech, employing a range of lexical, speaking style, and prosodic features in a boosting framework (2014, p. 236). In their article, "Recognition of stance strength and polarity in spontaneous speech", Levow et al. used three kinds of features to conduct their experiment with the ICSIboost (a boosting classifier): text-based features such as word unigram features, speaking style features such as spurt duration and the number of emphasized words, and prosodic features like pitch and intensity measures (2014, p. 238). Their experiment result showed that the word unigram feature alone achieves an accuracy of 80.5% on the stance detection task, which is the best performing feature compared to other features in this study.

In online discussions, users often support their statements with various arguments. However, unlike formal debates, arguments provided by online users are usually ambiguous, vague, implicit, or simply poorly worded (Boltužić and Šnajder, 2014, p. 50). Boltužić and Šnajder presented COMARG, a manually-annotated corpus for argument recognition of online discussions (2014). They also proposed a supervised model for "argument recognition based on the comment-argument comparison", which in contrast to stance detection, aims to understand the reasons underlying an opinion (Boltužić and Šnajder, 2014, pp. 50, 51).

For their argument recognition model, Boltužić and Šnajder employed three kinds of features: textual entailment (TE) features, semantic text similarity (STS) features, and stance alignment (SA) feature (2014). Textual entailment could help them determine whether the comment entails the argument phrase. While the semantic text similarity, as Agirre et al. describe, could measure "the degree of semantic equivalence between two texts" (as cited in Boltužić & Šnajder, 2014, p. 54). And the stance alignment feature is a binary feature whose value is one if a pro comment is paired with a pro argument. All three features are comment-argument comparison features, rather than extracted features from the comments or the arguments, which makes the model less domain-dependent. The result of their supervised classifier showed that the model with both TE and SA features performed the best.

Similar to the work that Boltužić and Šnajder have done, Somasundaran and Wiebe explored the utility of sentiment and arguing opinions for classifying stance in ideological

debates (n.d.). They used a total of six domains such as the existence of God, healthcare, and gun rights to conduct their experiment. Particularly, two kinds of features were exploited: arguing-based features like arguing-lexicon features and modal verb features, and sentiment-based features which are independent of arguing features (Somasundaran & Wiebe, n.d., p. 120). For experiments, Somasundaran and Wiebe also included a unigram system, which is based on unigram content information but no explicit opinion information (Somasundaran & Wiebe, n.d., p. 121). The experiment result indicated that the combination of arguing and sentiment features lead to the highest accuracy of 63.93% overall (Somasundaran & Wiebe, n.d., p. 122). In addition to that, the unigram system outperforms the sentiment feature system for all domains, showing that "what participants choose to speak about is a good indicator of ideological stance taking" (Somasundaran & Wiebe, n.d., p. 121).

Like Hasan and Ng described, online debaters could use emotional languages which may involve "sarcasm, insult, and questioning another debater's assumption and evidence", which makes them harder to study compared to parliament debates and company internal discussions (n.d., p. 816). In their article, "Extra-Linguistic Constraints on Stance Recognition in Ideological Debates", Hasan and Ng introduced two types of inter-post constraints on debate stance classification: user-interaction constraints (UC) and ideology constraints (IC) (n.d.). The user-interaction constraints were motivated by the observation that stance labels of the posts in a post sequence are not independent of each other, while the ideology constraints are only applicable to debate posts written by the same author in different domains (Hasan & Ng, n.d., pp. 817, 818). Hasan and Ng were able to create a

dataset by collecting debate posts of different topics from an online debate forum, which was later used in Sun et al. study (Sun et al., n.d.). The experiment result indicated that incorporating both UCs and ICs into the SVM stance classifier significantly improves the model performance, with an average improvement of 6.63% on accuracy (Hasan & Ng, n.d., p. 819).

2.2.2 Stance detection for fake news detection

In this section, relevant studies of stance detection on Fake News Challenge[1] and SemEval 2017 RumourEval[2]: Determining rumor veracity and support for rumors (Subtask A of SemEval 2017 Task 8) will be introduced (*RumourEval: Determining rumour veracity and support for rumours < SemEval-2017 Task 8*, n.d.).Although they are similar to the single Tweet stance classification since they studied the corpus of tweets, differences among them will also be presented.

As described by Kochkina et al., the Subtask A of SemEval 2017 Task 8 addresses the challenge of rumour stance classification (Kochkina et al., 2017, p. 475). This task is different from the single tweet stance classification since it addresses Twitter conversation threads. Each thread, according to Kochkina et al., includes a source tweet that initiates a conversation and associated nested tweets, which could be categorized into four stances: comment, support, deny and query tweets (2017, p. 476). After the pre-processing step, Kochkina et al. were able to extract a total of seven features such as

1 www.fakenewschallenge.org.html
2 http://alt.qcri.org/semeval2017/task8/

tweet lexicon, punctuation, relation to other tweets (2017, p. 477). Kochkina et al. proposed a novel branch-LSTM model with the Tree of Parzen Estimators (TPE) algorithm, which allowed them to process the whole branch of tweets, incorporate structural information of the conversation into the model, and achieve accuracy 0.784 on the testing set, outperforming all other systems in Subtask A (2017, pp. 477, 478).

The second best submission of Subtask A was achieved by Bahuleyan and Vechtomova, with an accuracy of 0.78 (2017, p. 461). In addition to tweet specific features such as punctuation that Kochkina et al. included in their model, Bahuleyan and Vechtomova came up with a hand-curated list of word features (cue features) and incorporated sentiment polarity score into their Gradient Boosting classifier (2017, pp. 462, 463). However, their experiment result indicated that including the sentiment score and similarity score into the classifier might not be helpful, and it is better if both cue features and tweet specific features were used in the model (Bahuleyan & Vechtomova, 2017, p. 463).

Similar to the impact of the rumourous tweet, the news industry also faces the challenge of false information, like Vosoughi et al. described, which might influence major events such as political elections (as cited in Mohtarami et al., 2018). Due to the tediousness of fact checking, automatic fact checking emerged and stance detection is one important step through the fact-checking process. Using the dataset provided by the Fake News Challenge, where each example contains a claim-document pair with four possible relationships (agree, disagree, discuss, unrelated), Mohtarami et al. presented an end-to-

end memory network for stance detection with the ability to extract snippets of evidence for the stance prediction (2018). The memory network model that Mohtarami et al. proposed was novel since it incorporated convolutional and recurrent neural networks, as well as a similarity matrix, which is the semantic similarity computed between claims and pieces of evidence (2018). The experiment result on the test data confirmed the importance of the proposed similarity matrix, which achieved an accuracy of 88.57% and was the best performing model in their evaluation (Mohtarami et al., 2018).

Another article in the Fake News Challenge that verifies the significance of similarity is presented by Riedel et al. (2018). As Riedel et al. describe, the stance detection system consists of "lexical and similarity features passed through a multi-layer perceptron (MLP) with one hidden layer" (2018). Specifically, they use only term frequency (TF) and term frequency-inverse document frequency (TF-IDF) for bag-of-words (BOW) representations of the text input, which eventually reached an accuracy of 88.46% (Riedel et al., 2018). This relatively straightforward system once again confirms the influence of similarity features in the task of stance detection, which could be considered as a baseline for the Fake News Challenge stance detection task.

2.2.3 Stance detection for tweets

Different from the Fake News Challenge and the Subtask A of SemEval 2017 Task 8, which address the stance detection task in conversational threads in tweets, researchers also studied how to recognize the stance of one single tweet in the recent years.

Prior to SemEval 2017, SemEval 2016 Task 6[3] already addresses the task of detecting stance in tweets (*Task 6: Detecting Stance in Tweets < SemEval-2016 Task 6*, n.d.). Mohammad et al. created the new stance dataset from tweets, organized the aforementioned shared task, and developed a linear-kernel SVM classifier (2017). Initially Mohammad et al. proposed five features including n-grams, sentiment lexicon, part-of-speech (POS) tag, and encodings which point out the presence/absence of positive and negative emotions (2017). Additionally, they used a specific target feature to indicate the presence/absence of the target of interests in the tweet (Mohammad et al., 2017). And their experiment result showed that while the sentiment lexicon alone is not sufficient, adding the target features could lead to small improvements for the model performance, achieving an F-score of 70.3 (Mohammad et al., 2017).

The best performing system for SemEval 2016 Task 6, with an average F-score of 67.8, was achieved by Zarrella and Marsh (2016). They employed a recurrent neural network (RNN) with features learned via two large unlabeled datasets, and they also trained embeddings of words with the word2vec skip-gram method, which was later used to learn sentence representation through a hashtag prediction auxiliary task (Zarrella & Marsh, 2016). Their experiment result showed that the majority class of their dataset significantly outperformed the corresponding minority class.

Another article that exploits the dataset of SemEval 2016 Task 6 is done by Sun et al. In addition to the aforementioned corpus, they also used the dataset that was collected by

3 http://alt.qcri.org/semeval2016/task6/

Hasan and Ng, which is from an online debate forum (as cited in Sun et al., n.d., p. 2404). It is noticeable that the average lengths of these two corpora are drastically different (114 for Hasan & Ng and 18 for SemEval), which enables them to evaluate the performance of their proposed model in different settings (Sun et al., n.d., p. 2404). Sun et al. presented a hierarchical neural attention model (HAN), which contains a linguistic attention part and a hyper attention part (n.d., p. 2400). The linguistic attention is able to "learn the correlations between document representation and different linguistic feature sets", while the hyper attention is able to "adjust the weight of different feature sets" to achieve the best result (Sun et al., n.d., p. 2401). A standard LSTM model was employed in the process to learn the document, sentiment, dependency, and argument representation (Sun et al., n.d.). In the experiment result, the proposed model (HAN) outperformed several baseline systems including SVM and LSTM and confirmed the effectiveness of the target information and argument information (Sun et al., n.d., p. 2406).

Besides the SVM classifiers and the Long Short-Term Memory (LSTM) model, another model (TAN) that was used in Sun et al. as a baseline system is proposed by Du et al. Similar to the HAN model, the TAN model also is a neural network-based model, which incorporates target-specific information by an attention mechanism (Du et al., n.d.). This Target-specific Attentional Network (TAN) model combines the recurrent neural network (RNN) with long-short memory (LSTM) and target-specific attention extractor (Du et al., n.d.). To evaluate the model performance, Du et al. conducted the experiments on two datasets, one English and one Chinese, and the TAN model outperforms all baselines

significantly, which demonstrates that TAN is a language-independent model across different languages (n.d.).

2.3 Models categorization/comparison across various domains

Previous works on the stance detection task mainly exploited three types of models. The first one is the support vector machine (SVM) model with various methods of feature extraction (Boltužić and Šnajder, 2014; Hasan & Ng, n.d.; Mohammad et al., 2017; Somasundaran & Wiebe, n.d.). The second type of model that was exploited is the gradient boosting algorithm (Bahuleyan & Vechtomova, 2017; Levow et al., 2014). The last category is the neural network model, which was later developed into various approaches mentioned before (Du et al., n.d.; Mohtarami et al., 2018; Sun et al., n.d.; Zarrella & Marsh, 2016). Although the neural network model is the most complicated one among these three categories, the straightforward model of SVM sometimes also achieves satisfying evaluation results on the test data.

For this study, it focused on detecting the stance of a single tweet. Several research questions, listed below, were studied and experimented.

- Research Question 1: Which machine learning algorithm works the best for the task of stance detection?
- Research Question 2: Is sentiment helpful for predicting stance?
- Research Question 3: Is the target information valuable for classifying stance?

- Research Question 4: Can features, which are extracted from one target, be generalized to other targets?

3 Methodology

In the section, a detailed plan of research and experiment for stance detection on tweets will be presented. The following subsections will:

- describe the dataset used in this research,
- explain the process of data aggregation,
- demonstrate the properties of the dataset
- illustrate the research method and evaluation metric,
- elaborate on the anticipated implication of the research.

3.1 Data

3.1.1 Initial dataset

The initial dataset[4] comes from the SemEval 2016 Task 6, organized by Mohammad et al. (2017). Specifically, more than 4,000 tweets were collected and annotated for whether one can detect positive or negative stance towards one of the five topics: “Atheism”, “Climate Change is a Real Concern” (“Climate”), “Feminist Movement” (“Feminist”), “Hillary Clinton”, and “Legalization of Abortion” (“Abortion”). Task 6 contains two subtasks - subtask A for supervised learning and subtask B for unsupervised learning. In this study, the dataset of subtask A was used since the target provided in the testing set is

4 http://alt.qcri.org/semeval2016/task6/data/uploads/stancedataset.zip

also be in the training set. The dataset of subtask B only contains instances towards one target “Donald Trump”, which will be used for testing and evaluation purposes.

The original dataset is stored in CSV format. For each tweet, it has five columns: the first column is the textual content of the tweet; the second column is the associated target within the five categories; the third column is the stance of this particular tweet, which falls in three types - Favor, Against, and None; the fourth column is the indicator of whether the target of opinion in the tweet is the same as the given target of interest; the last column is the sentiment label of the tweet, which also fall into three categories - positive, negative, and other. The last three columns were results of annotation from CrowdFlower[5] with the questionnaire provided by Mohammad et al. (2017).

One example of the dataset is shown below:

Text: *Use your brain, keep Hillary out of the White House.Clinton2016*

Target: Hillary Clinton

Stance: AGAINST

Opinion towards: 1. The tweet explicitly expresses opinion about the target, a part of the target, or an aspect of the target.

Sentiment: neg

3.1.2 Data aggregation and separation

The initial dataset was separated into training and testing sets. Like mentioned before, both training and testing sets have five targets. In order to conduct this study, the training

5 http://www.crowdflower.com

and testing sets were first merged into a complete one, which includes 4,163 tweets. Then this aggregated dataset was separated into five subsets depending on the five targets.

3.1.3 Properties of the stance dataset

The distribution of the aggregated dataset based on stance is shown in Table 1. Notice that except for tweets that are related to the target "Climate Change is a Real Concern" ("Climate"), all other tweets are predominantly against their expressed target, which weigh more than 50% in each of its own categories. Moreover, the percentage of each target subset was also an indicator for the expected accuracy of predicting the majority category. For instance, regarding the "Feminist" subset, its baseline accuracy will be 53.85%.

Table 1: Distribution of instances based on stance

	% of instances			
Target	**Favor**	**Against**	**None**	**Total #**
Atheism	16.92%	63.30%	19.78%	733
Climate	59.40%	4.61%	35.99%	564
Feminist	28.24%	53.85%	17.91%	949
Hillary Clinton	16.57%	57.42%	26.02%	984
Abortion	17.90%	58.31%	23.79%	933
Donald Trump	20.93%	42.29%	36.78%	707
Total	24.74%	49.47%	25.79%	4870

Table 2: Distribution of instances based on sentiment

	% of instances			
Target	**Positive**	**Negative**	**Other**	**Total #**
Atheism	60.03%	35.20%	4.77%	733
Climate	31.03%	50.18%	18.79%	564
Feminist	18.34%	76.92%	4.74%	949
Hillary Clinton	30.18%	65.85%	3.96%	984
Abortion	26.26%	67.95%	5.79%	933
Total	31.97%	61.33%	6.70%	4163

Table 2 illustrates the distribution of instances based on sentiment. Observe that tweets corresponding to all targets, except for “Atheism”, are mainly expressed negative sentiment. Additionally, tweets regarding the target “Feminist Movement” (“Feminist”) have the highest polarity towards the negative sentiment (76.92%).

Table 3: Percentage distribution of instances in the tweet dataset

	Opinion towards			
Target	**Target**	**Other**	**No one**	**Total #**
Atheism	49.25%	46.38%	4.37%	733
Climate	60.82%	30.50%	8.69%	564
Feminist	68.28%	27.40%	4.32%	949
Hillary Clinton	60.37%	35.06%	4.57%	984
Abortion	63.67%	30.98%	5.36%	933
Total	61.01%	33.77%	5.21%	4163

Table 3 shows the percentage distribution of instances regarding whether the opinion is expressed towards the given target or not. It also indicates that although opinions of the dataset are mainly expressed towards the given target (61.01% in total), the percentage of opinion towards others varies across targets from 27.4% to 46.38%.

Table 4: Percentage distribution of instances by target of opinion

	Opinion towards			
Stance	**Target**	**Other**	**No one**	**Total #**
Favor	94.23%	5.11%	0.66%	1057
Against	72.75%	26.54%	0.71%	2110

Table 4 demonstrates the distribution of tweets by opinion for the "favor" and "against" stance labels despite given targets. It is noticeable that for tweets with the annotated unfavorable stance, 26.54% of the opinions are expressed towards something or someone else rather than the target. This happens since for a number of tweets, the target is not explicitly mentioned in the text, but the annotators determine the stance towards the target. Like the example mentioned in the introduction, within the text "We live in a world where people care more about religious rights than consent & reproductive rights....", it didn't mention any terms such as "abortion" or "pro-life", and yet expressed a favorable stance towards the topic of "Legalization of Abortion".

3.2 Research method and evaluation metric

In this study, LightSIDE[6] was the primary tool to conduct experiments and run tests. It is a free and open text-mining toolkit developed by Elijah Mayfield at Carnegie Mellon University. LightSIDE offers a straightforward GUI environment for users to easily extract text features, run machine learning and text mining algorithms, and conduct error analysis. Following sections first illustrated the feature extraction process in LightSIDE, then explained various approaches to conduct experiments regarding aforementioned research questions and covered the evaluation metrics.

6 http://ankara.lti.cs.cmu.edu/side/download.html

3.2.1 Feature selection

LightSIDE provides numerous features for the user to select in the feature extraction process. Some features are selected to construct the feature table across this experiment. Prior work for sentiment analysis showed that the most practical features are n-grams and sentiment lexicons, while others like negation features, part-of-speech features, and punctuation might have a smaller effect (Kiritchenko et al., 2014; Mohammad et al., 2013, 2017, p. 10; Nakov et al., 2019; Rosenthal et al., 2019). Since the text length of a tweet is limited, textual information is valuable for feature extraction. In addition to basic unigrams, bigrams and trigrams are also used for feature selection as they will be able to remember word orders and represent phrases or collocations of words that often appear together.

Because of the short length of tweets, punctuations such as an exclamation mark play an important part when people express their stances towards a certain topic. After manual speculation, the tweet dataset is relatively clean and organized and there is no need to exclude punctuations. Two particular features are included in the feature extraction step to reduce the size of feature table and gain generality. One is stemming, which aims to reduce words to their base form. Words like “run”, “running”, and “ran” will all count as the same concept. With stemming, these words will be represented by a single “run” feature, dramatically reducing the size of the feature table while losing inflection. The other feature to gain generality is to skip stopwords. In LightSIDE, it has a list of 118 common words such as “and” or “the”, which don’t carry actual meaning of the content

but serve as function words to connect text together. All these words were excluded in order to reduce the size of the feature table.

Additionally, LightSIDE offers the POS (part-of speech tags) n-grams feature, serving as a proxy for complicated syntactic structures. LightSIDE's part of speech is based on computational linguistics research and employed the Stanford POS tagger, which is able to identify more than thirty possibilities such as "VBP" (a non-third-person singular verb in the present tense) or "PRP" (a personal pronoun, such as "he" or "we") (Mayfield et al., n.d.). The POS n-grams features were also evaluated for the experiment in the latter process.

3.2.2 Methods for research questions

Regarding the four research questions introduced at the end of the literature review section, several steps were followed to conduct the experiment and analyze the result.

First of all, to test and evaluate the performances of various machine learning algorithms (RQ1), LightSIDE offers five algorithms for the users: Naive Bayes, Logistic Regression, Linear Regression, Support Vector Machines, and Decision Trees. However, since the Linear Regression algorithm cannot predict nominal class values, in this case, the target class of stance. Ergo Linear Regression was not used in this study. This study first used the five target-specific subsets to do feature extraction, then built and tested the four algorithms except Linear Regression. These four algorithms were also be tested on the

dataset as a whole. Based on which algorithm performed the best for the whole dataset and previous studies, one algorithm was used for all the following experiments.

So as to test whether sentiment polarity is valuable for stance detection (RQ2), during the feature extraction process, column features option of "sentiment" in the CSV file was selected in LightSIDE, which was able to provide additional information in addition to the tweet text. Then procedures similar to RQ1 were followed to test and evaluate the effect of adding sentiment information to the feature table and the machine learning model. Analogous steps were also taken for RQ3, with only adding the "target" as a column features option into the feature selection procedure.

Like mentioned in the feature selection section, to see whether the POS n-grams feature is valuable for the stance detection task, this feature was also tested in LightSIDE to see its impact. In addition to using the tweet text with the sentiment polarity or the target information to construct the feature table, one approach taken both sentiment polarity and the target information was experimented. Furthermore, to experiment if the tweet solely is sufficient for the stance detection task, another approach using only the tweet text with predicted sentiment and target was tested. First, the tweet text was used to predict the sentiment polarity and the target information. Then these three columns jointly predicted the stance of the tweet. Based on the performance of different approaches, one was selected to test for generality for Research Question 4.

In order to test if features from one subset could be generalized to other datasets (RQ4), features from one particular subset was extracted, which was later used for model building and testing on other four subsets and the aggregated dataset. In addition, features were also tested on the whole dataset and new dataset with target "Donald Trump" to see its generality.

3.2.3 Evaluation metric

To evaluate the performance of various models and different representations of feature tables, LightSIDE provides two statistical metrics for users to judge a model. The first one is accuracy as percentage of correctly predicted labels. It is a straightforward classification metric to understand and interpret. The accuracy measure was used as an evaluation metric through this study. Formula for calculating accuracy is listed below.

- $Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions}$(Bhatia, 2018)

For validation purposes, LightSIDE offers cross-validation as well as supplied test sets option, which allows the user to import a new test set for the training model. For this study, models regarding Research Question 2 and afterwards were evaluated on 10-fold random cross-validation and a new test set provided by SemEval 2016 Task 6 subtask B, which only includes instances towards the target "Donald Trump".

3.3 Implication of research

Stimulated by the growth of use in microblog platforms such as Twitter, companies and media organizations are increasingly seeking options to analyze people's opinion and

stance towards their product and services (Du et al., n.d.). Given the experiments conducted in this study, it would provide a preliminary insight towards the topic of stance detection, specifically on the Twitter dataset. We would able to understand the reasons and arguments underpinning Twitter users' opinions. Understanding the reasons and arguments has various benefits and applications, ranging from brand analysis to political related research (Boltužić & Šnajder, 2014).For instance, detecting the stance of a Twitter user regarding a rumourous information would provide an indirect way to identify potential rumors (Bahuleyan & Vechtomova, 2017).

Identifying viewpoints from such short text like a tweet would provide explanations on how people express themselves and how they choose words and languages. Being able to detect stance automatically is very helpful when dealing with public resonance and associated rumours, as misinformation spread on social media has potential negative impact on the related situation. (Bahuleyan & Vechtomova, 2017; Zhao et al., 2015). For the stance detection task particularly, this study would test on various approaches and offer insights on whether the target information and the sentiment polarity would be helpful, which might be beneficial for future research.

4 Results and Discussion

Results of aforementioned experiments are presented in this section. First, performance of different algorithms was illustrated. Then different approaches including various features were illustrated and compared. Finally, the generality issue of applying target-specific features to other datasets was addressed, followed by a discussion section to analyze the results.

4.1 Performance of algorithms

The performance of various machine learning algorithms is presented in Table 5. There are four algorithms in LightSIDE that could be used: Naive Bayes, Logistic Regression, Support Vector Machine, and Decision Trees. All of these models used tweet-only features. For the five target-specific subsets, Naive Bayes and Logistic Regression outperformed the other two algorithms. Although the accuracy of the Logistic Regression model is slightly higher than that of the Naive Bayes model for the subsets “Atheism”, “Climate”, and “Feminist”, statistically there is no difference between the performance of these two algorithms. However, for the aggregated dataset, the Naive Bayes model performed marginally better than the Logistic Regression model, which is the reason that all following experiments are based on the Naive Bayes algorithm.

Table 5: Performance of different algorithms

Target/Feature from	Baseline Accuracy	Different algorithms: Naive Bayes	Logistic Regression	Support Vector Machines	Decision Trees
Atheism	0.6330	0.6876	**0.6999**	0.6617	0.6412
Climate	0.5940	0.6915	**0.7021**	0.6720	0.6755
Feminist	0.5385	0.5690	**0.5753**	0.5385	0.5332
Hillary Clinton	0.5742	**0.6535**	0.6413	0.6220	0.5955
Abortion	0.5831	**0.6731**	0.6731	0.6517	0.5949
Total	0.4947	**0.6166**	0.6049	0.5693	-

It is noticeable that for the aggregated dataset, the best model, Naive Bayes algorithm, achieved an accuracy of 0.6166, which is not high for a classification task. Particularly, the subset of "Feminist" performed worse than any other subsets, with an best accuracy of 0.5733. The subset of "Climate" reached the highest accuracy of 0.7021 with the Logistic Regression algorithm. Additionally, in terms of training and testing time, the Decision Trees algorithm took much longer time than other algorithms. And this algorithm failed to receive a result for the aggregated dataset, perhaps due to the larger feature table and more instances compared to the subsets.

4.2 Performance of models with different approaches

Table 6 illustrated the results of different feature approaches with cross validation as the testing method. For the five target-specific subsets, there are a total of five approaches trained and tested: only tweet textual feature, text and POS n-grams feature, text and sentiment polarity feature, text and target information feature, and feature of text with both sentiment polarity and target information. For the aggregated dataset, one additional approach with predicted sentiment and target features is also trained and tested.

Table 6: Performance of different feature extraction approaches on cross-validation method

Approaches	Target/Feature from Atheism	Climate	Feminist	Hillary Clinton	Abortion	Total
Baseline Accuracy	0.6330	0.5940	0.5385	0.5742	0.5831	0.4947
Only text	0.6876	0.6915	0.5690	0.6535	0.6731	0.6166
Text + POS	0.6712	0.6684	0.5490	0.6240	0.6227	0.5904
Text + Sentiment	**0.7135**	0.6773	**0.5954**	0.6728	**0.6763**	0.6169
Text + Target	0.6876	0.6915	0.5690	0.6535	0.6731	0.6272
Text + Sentiment&Target	0.7026	**0.6950**	0.5827	**0.6850**	0.6763	**0.6284**
Text + Predict Sentiment&Target	-	-	-	-	-	0.6221

For all datasets, adding the POS n-grams feature didn't improve the performance of the model. The accuracy of the model based on the aggregated dataset with only textual features (0.6166) is significantly higher than the accuracy of the model with textual and POS n-grams features (0.5904). Different from the performance of models with text and POS n-grams features, adding the sentiment polarity as a column feature into the model improved all models' performance except for the subset of "Climate". Specifically, for the subset of "Atheism", the model with sentiment polarity achieved an accuracy of 0.7135, which is the highest number in the test. However, in the case of the aggregated dataset, including the sentiment polarity didn't improve the model performance so much, compared to the model with only textual features.

For the experiment result of including the target information into the feature extraction process, all five subsets performed the same compared to the performance with only textual features. This is because that for these subsets, including the target information only adds one more feature into the table, which didn't result in a large impact to the model. However, for the aggregated dataset, the model with target information

statistically improved the accuracy (0.6272) when comparing to the model with only textual features (0.6166).

Another approach that was tested in the study was to include both sentiment polarity and the target information as column features. And adding both of them into the feature table improved the model performance for all five subsets and the aggregated dataset. For the subset “Climate” and the subset “Hillary Clinton”, they achieved their highest accuracies in the test with an accuracy of 0.6950 for the subset “Climate” and an accuracy of 0.6850 for the subset “Hillary Clinton”. For the aggregated dataset, it achieved an accuracy of 0.6284, which is a significant improvement compared to the model with only textual features. However, compared to the model with the target information with an accuracy of 0.6272, adding both sentiment polarity and target information didn’t significantly improve the model performance. And adding both of them is a marginal improvement compared to the model with sentiment polarity.

The last approach included in this study is to generate a feature table with predicted sentiment polarity and target information, then jointly detect the stance. Since only the aggregated dataset has multiple targets, it could be used for this particular approach. And this method achieved an accuracy of 0.6221, which is higher than the accuracy of the method with only textual features.

In addition to the cross validation as the testing method, aforesaid approaches (textual features, POS n-grams features, sentiment polarity features, and target information

features) were also tested on a new subset with the target “Donald Trump”. The result is shown in Table 7. The baseline accuracy for this subset is the percentage distribution of the majority category, which is 0.4229 shown in Table 1. It is noticeable that all accuracies of testing on this new target dataset are significantly lower than the results on cross validation method. The best target-specific feature for detecting the stance regarding the target “Donald Trump” is from the “Abortion” subset, resulting in an accuracy of 0.4031. Also, the aggregated dataset didn’t do well on detecting the stance towards the “Donald Trump” dataset. The best performance for the aggregated dataset leads to an accuracy of 0.3663, which is worse than many of the results from the target-specific subsets. Another pattern for detecting the stance of the “Donald Trump” dataset is that most of the best performing results came from the combination of both text and sentiment polarity features.

Table 7: Performance of feature extraction approaches testing on “Donald Trump” dataset

	Target/Feature from					
Approaches	**Atheism**	**Climate**	**Feminist**	**Hillary Clinton**	**Abortion**	**Total**
Only text	**0.3805**	**0.3451**	0.3861	0.3479	0.3946	0.3607
Text + POS	0.3777	0.3027	0.3423	0.3678	0.3876	0.3437
Text + Sentiment	0.3437	0.3253	**0.4017**	**0.3748**	**0.4031**	**0.3663**
Text + Target	-	-	-	-	-	0.3267

4.3 Performance of models with target-specific features

In order to test whether features from one subset could be generalized to other datasets, experiments of testing target-specific features to other subsets were conducted. The

results are shown in Table 8. In addition to testing on subsets, all five target-specific features are also tested on the aggregated dataset and the new "Donald Trump" subset. As shown in Table 8, the subsets "Atheism" and "Climate" didn't get high accuracies on generalizing their target-specific features to other subsets, while the other three subsets did a relatively better job. The highest accuracy was achieved by applying the "Feminist" feature table to the "Hillary Clinton" subset, with an accuracy of 0.4543. Regarding the performance on the aggregated dataset and the "Donald Trump" subset, feature table from the "Abortion" subset did better than other subsets, resulting in an accuracy of 0.4879 on the aggregated dataset and an accuracy of 0.4031 on the "Donald Trump" subset.

Table 8: Performance of generalizing target-specific features to datasets with different targets

	Test Set						
Feature from	**Atheism**	**Climate**	**Feminist**	**Hillary Clinton**	**Abortion**	**Total**	**DT**
Baseline Accuracy	0.6330	0.5940	0.5385	0.5742	0.5831	0.4947	0.4229
Atheism	-	0.3032	0.2561	0.3059	**0.3655**	0.4014	0.3437
Climate	0.1951	-	0.2276	**0.2459**	0.2347	0.3024	0.3253
Feminist	0.2974	0.328	-	**0.4543**	0.4309	0.4797	0.4017
Hillary Clinton	0.2551	0.2535	0.412	-	**0.4212**	0.4629	0.3748
Abortion	**0.4434**	0.273	0.4057	0.4238	-	**0.4879**	**0.4031**

4.4 Discussion

First of all, different from the SVM classifier that Mohammad et al. employed, this study indicated that a simple Naive Bayes model could achieve a better performance. This may be due to the short length of the tweet dataset, which has a character limit on each tweet. Like Wang and Manning mentioned, that for short snippets, Naive Bayes performs better

than SVM, while for longer documents the opposite result holds (n.d.). When choosing the appropriate machine learning algorithm to perform text classification tasks, the characteristics of the dataset should also be brought into consideration since it may also influence the performance of the machine learning model.

Secondly, including the sentiment polarity information into the feature table can improve the performance regarding the stance detection task. As shown in Table 6, accuracies of most subsets based on cross validation method increased except for the subset “Climate”. In Table 7, outcomes also indicated that adding the sentiment polarity into the feature table could improve the performance for detecting the stance towards a brand-new dataset. However, the accuracy for the aggregated dataset didn’t improve much, which coincides with the opinion from Mohammad et al. that sentiment alone is not sufficient for the stance detection task (2017, p. 13).

Thirdly, adding the target information into the feature extraction process did improve the performance of the stance detection task. And including both the sentiment polarity and the target information also improved the accuracy of the stance detection result. However, although all improvements are statistically significant, the actual increase in the accuracy number is relatively small. For adding the target information, the accuracy of detecting stance for the aggregated dataset increased from 0.6166 to 0.6272. For adding both sentiment polarity and target information, the accuracy went up from 0.6166 to 0.6284, which deviated from the observation in experiments of Mohammad et al. that

combination of features like “n-grams + target + sentiment” didn’t improve the performance (2017, p. 13).

Lastly, target-specific features performed poorly on detecting the stance towards other target-specific subsets. None of these features achieved an accuracy of more than 0.5, with the highest accuracy of 0.4543 for applying features from the “Feminist” subset to the “Hillary Clinton” subset. It is noticeable that features from one subset resulted in a higher accuracy when applying to a more similar subset. For instance, as shown in Table 8 regarding the features from the “Atheism” subset, the experiment on the “Abortion” subset reached an accuracy of 0.3655, which is higher than any other subsets. After calculation, the “Atheism” subset has a cosine similarity of 0.9261 with the “Abortion” subset, which is also higher than the similarity between the “Atheism” subset with any other subsets. The same pattern could also be applied to the features from the “Hillary Clinton” subset. Although the feature table did unsatisfactorily on all subsets, applying it to the “Abortion” subset achieved a relatively high accuracy compared to the results of other subsets. The cosine similarity (0.9179) between the “Hillary Clinton” subset and the “Abortion” subset is a higher figure compared to that between the “Climate” subset with other subsets. Moreover, the last three subsets achieved a better result when applying to the aggregated dataset, which after manual inspection, might because of their larger number of instances in the dataset compared to the other two. And the larger number of instances might also be the reason why they performed better on the new “Donald Trump” subset.

5 Conclusion

This study first researches the previous work on the topic of stance detection on several different fields such as online forum discussion, application on fake news detection, and stance detection on tweet, which is directly related to this study. Then this study proposed several different experiment approaches to test whether adding the sentiment polarity and target information is beneficial for the stance detection task. And it showed that adding both features into the machine learning model could improve the experiment results, however, neither the sentiment polarity nor the target information alone is not sufficient enough. The textual information is the core of such stance detection task. Additionally, the experiment result in this study showed that it is hard to generalize target-specific features to a dataset with different targets. And the similarity between datasets might have an impact on applying target-specific features to another dataset.

Future study could be in several directions. Regarding the results of applying target-specific features to different datasets, one possible direction is to analyze the relationship between such results and the similarities among datasets with different targets. People might tend to use similar language when they expressed their opinion about related topics. Another direction might be to design an automatic system to detect the target of a certain document, no matter if it is a length document or a short tweet. This would further help this stance detection task to analyze the impact of including the target information.

There are several limitations of this work. First, in this study, all target information is acquired through manual annotation, therefore for each subset, they only had one target column, which is not possible to generalize to other subsets. Secondly, the data size in this study might not be sufficient as there is only a total of 4,163 tweets. Adding more data might be helpful for extracting enough general features and applying them to datasets with different targets, which might lead to a better generalization result. Lastly, due to the time limit, this study didn't perform sophisticated experiments such as neural network models. There might be other machine learning algorithms that perform better on the tweet dataset.

6 References

Bahuleyan, H., & Vechtomova, O. (2017). UWaterloo at SemEval-2017 Task 8: Detecting Stance towards Rumours with Topic Independent Features. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 461–464. https://doi.org/10.18653/v1/S17-2080

Bhatia, A. (2018, June 25). *Accuracy will not give the correct picture always…..* Medium. https://medium.com/@ab9.bhatia/the-simplest-model-evaluation-metric-for-classification-models-is-accuracy-it-is-the-percentage-of-75290a9aa126

Boltužić, F., & Šnajder, J. (2014). Back up your Stance: Recognizing Arguments in Online Discussions. *Proceedings of the First Workshop on Argumentation Mining*, 49–58. https://doi.org/10.3115/v1/W14-2107

Cabrio, E., & Villata, S. (2018). Five Years of Argument Mining: A Data-driven Analysis. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 5427–5433. https://doi.org/10.24963/ijcai.2018/766

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Du, J., Xu, R., He, Y., & Gui, L. (n.d.). *Stance Classification with Target-Specific Neural Attention Networks*. 7.

Epiville: How to Calculate Kappa. (n.d.). Retrieved April 13, 2020, from

https://epiville.ccnmtl.columbia.edu/popup/how_to_calculate_kappa.html

Fake News Challenge. (n.d.). Retrieved April 6, 2020, from http://www.fakenewschallenge.org/

Hasan, K. S., & Ng, V. (n.d.). *Extra-Linguistic Constraints on Stance Recognition in Ideological Debates*. 6.

Kiritchenko, S., Zhu, X., & Mohammad, S. M. (2014). Sentiment Analysis of Short Informal Texts. *Journal of Artificial Intelligence Research*, *50*, 723–762. https://doi.org/10.1613/jair.4272

Kochkina, E., Liakata, M., & Augenstein, I. (2017). Turing at SemEval-2017 Task 8: Sequential Approach to Rumour Stance Classification with Branch-LSTM. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 475–480. https://doi.org/10.18653/v1/S17-2083

Levow, G.-A., Freeman, V., Hrynkevich, A., Ostendorf, M., Wright, R., Chan, J., Luan, Y., & Tran, T. (2014). Recognition of stance strength and polarity in spontaneous speech. *2014 IEEE Spoken Language Technology Workshop (SLT)*, 236–241. https://doi.org/10.1109/SLT.2014.7078580

Lippi, M., & Torroni, P. (2016, March 5). Argument Mining from Speech: Detecting Claims in Political Debates. *Thirtieth AAAI Conference on Artificial Intelligence*. Thirtieth AAAI Conference on Artificial Intelligence. https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12164

Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1–167.

https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Mayfield, E., Adamson, D., & Rosé, C. P. (n.d.). *Researcher's Workbench User Manual*. 55.

Mohammad, S. M., Kiritchenko, S., & Zhu, X. (2013). NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. *ArXiv:1308.6242 [Cs]*. http://arxiv.org/abs/1308.6242

Mohammad, S. M., Sobhani, P., & Kiritchenko, S. (2017). Stance and Sentiment in Tweets. *ACM Transactions on Internet Technology*, *17*(3), 1–23. https://doi.org/10.1145/3003433

Mohtarami, M., Baly, R., Glass, J., Nakov, P., Marquez, L., & Moschitti, A. (2018). Automatic Stance Detection Using End-to-End Memory Networks. *ArXiv:1804.07581 [Cs]*. http://arxiv.org/abs/1804.07581

Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., & Stoyanov, V. (2019). SemEval-2016 Task 4: Sentiment Analysis in Twitter. *ArXiv:1912.01973 [Cs]*. http://arxiv.org/abs/1912.01973

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1–135. https://doi.org/10.1561/1500000011

Patodkar, V. N., & I.R, S. (2016). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *IJARCCE*, *5*(12), 320–322. https://doi.org/10.17148/IJARCCE.2016.51274

Riedel, B., Augenstein, I., Spithourakis, G. P., & Riedel, S. (2018). A simple but tough-to-beat baseline for the Fake News Challenge stance detection task.

ArXiv:1707.03264 [Cs]. http://arxiv.org/abs/1707.03264

Rory, D., Katarzyna, B., & Chris, R. (2016). Mining Ethos in Political Debate. *Frontiers in Artificial Intelligence and Applications*, 299–310. https://doi.org/10.3233/978-1-61499-686-6-299

Rosenthal, S., Mohammad, S. M., Nakov, P., Ritter, A., Kiritchenko, S., & Stoyanov, V. (2019). SemEval-2015 Task 10: Sentiment Analysis in Twitter. *ArXiv:1912.02387 [Cs]*. http://arxiv.org/abs/1912.02387

RumourEval: Determining rumour veracity and support for rumours < SemEval-2017 Task 8. (n.d.). Retrieved April 6, 2020, from http://alt.qcri.org/semeval2017/task8/

Somasundaran, S., & Wiebe, J. (n.d.). *Recognizing stances in ideological on-line debates*. 9.

Strijbos, J.-W., Martens, R. L., Prins, F. J., & Jochems, W. M. G. (2006). Content analysis: What are they talking about? *Computers & Education*, *46*(1), 29–48. https://doi.org/10.1016/j.compedu.2005.04.002

Sun, Q., Wang, Z., Zhu, Q., & Zhou, G. (n.d.). *Stance Detection with Hierarchical Attention Network*. 11.

Task 6: Detecting Stance in Tweets < SemEval-2016 Task 6. (n.d.). Retrieved April 6, 2020, from http://alt.qcri.org/semeval2016/task6/

Viera, A. J., & Garrett, J. M. (n.d.). Understanding Interobserver Agreement: The Kappa Statistic. *Family Medicine*, 4.

Wang, S., & Manning, C. D. (n.d.). *Baselines and Bigrams: Simple, Good Sentiment and Topic Classification*. 5.

Zarrella, G., & Marsh, A. (2016). MITRE at SemEval-2016 Task 6: Transfer Learning

for Stance Detection. *ArXiv:1606.03784 [Cs]*. http://arxiv.org/abs/1606.03784

Zhao, Z., Resnick, P., & Mei, Q. (2015). Enquiring Minds: Early Detection of Rumors in Social Media from Enquiry Posts. *Proceedings of the 24th International Conference on World Wide Web - WWW '15*, 1395–1405. https://doi.org/10.1145/2736277.2741637