UWB at SemEval-2016 Task 6: Stance Detection

Peter Krejzl
University of West Bohemia
Faculty of Applied Sciences
Dept. of Computer Sci. and Eng.
Univerzitni 8, 30614 Pilsen
Czech Republic
pkrejzl@gmail.com

Josef Steinberger
University of West Bohemia
Faculty of Applied Sciences
NTIS Centre
Univerzitni 8, 30614 Pilsen
Czech Republic
jstein@kiv.zcu.cz

Abstract

This paper describes our system participating in the SemEval 2016 task: Detecting stance in Tweets. The goal was to identify whether the author of a tweet is in favor of the given target or against. Our approach is based on a maximum entropy classifier, which uses surface-level, sentiment and domain-specific features. We participated in both the supervised and weakly supervised subtasks and received promising results for most of the targets.

1 Introduction

Stance detection has been defined as automatically detecting whether the author of a piece of text is in favor of the given target or against it. In the third class, there are the cases, in which neither inference is likely. It can be viewed as a subtask of opinion mining and it stands next to the sentiment analysis. The significant difference is that in the case of sentiment analysis, systems determine whether a piece of text is positive, negative, or neutral. However, in stance detection, systems are to determine the author's favorability towards a given target and the target even may not be explicitly mentioned in the text. Moreover, the text may express positive opinion about an entity contained in the text, but one can also infer that the author is against the defined target (an entity or a topic). This makes the task more difficult, compared to the sentiment analysis, but it can often bring complementary information.

There are many applications which could benefit from the automatic stance detection, including information retrieval, textual entailment, or text summarization, in particular opinion summarization. Twitter was selected as the source of the text because of its popularity and because people express stance implicitly or explicitly there.

We first shortly introduce the task in Section 2 and the available dataset.[1] In Section 3, we describe our preprocessing, the implemented approach and system's features. It is followed by the setup for each analysed topic and a discussion of official results (Section 4).

2 Task Description

The **Detecting Stance in Tweets** task[2] (Mohammad et al., 2016) had two independent subtasks: supervised and weakly supervised stance identification.

The supervised task (subtask A) tested stance towards five targets: *Atheism*, *Climate Change is a Real Concern*, *Feminist Movement*, *Hillary Clinton*, and *Legalization of Abortion*. Participants were provided 2814 labeled training tweets for the five targets. An example tweet annotated as IN FAVOR: *These pics of #pornstars with/without makeup? Just perpetuating the myth that women need makeup to be considered pretty.* (the *Feminist Movement* target, ID: 1017).

A detailed distribution of stances for each target is given in Table 1. The distribution is not uniform and there is always a preference towards a certain stance (e.g., 59% tweets about *Atheism* are labelled as AGAINST).

[1] Details can be found in the overview paper (Mohammad et al., 2016).

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

Proceedings of SemEval-2016, pages 408–412,
San Diego, California, June 16-17, 2016.

It naturally reflects the real-world scenario, in which a majority of people tend to one of the stances. This is also depending on the source of the data. For example, in the case of *Legalization of Abortion*, we can assume that the distribution will be significantly different in religious communities than in atheistic communities.

For the weakly supervised task (subtask B), there were no labelled training data but participants could use a large number of tweets related to the single target: *Donald Trump*. Example: *There are so many reasons to dislike #HillaryClinton but Half Human and Half Orangutan #DonaldTrump takes it to next level.* (ID: 589371241204711424).

Due to Twitter legal requirements, the dataset for subtask B contained only tweet ids and participants had to download those tweets using the provided script. The dataset contains 78256 tweet ids generated by searching for the tags #DonaldTrump and #trump2016. Unfortunately, some of those tweets did not exist at the download time, as their authors removed them. Our dataset contained 69454 tweets.

Topic	FAV	AG	NONE	TOT
Atheism	92 (18%)	304 (59%)	117 (23%)	513
Climate Change is a Real Concern	212 (54%)	15 (4%)	168 (43%)	395
Feminist Movement	210 (32%)	328 (49%)	126 (19%)	664
Hillary Clinton	112 (18%)	361 (56%)	166 (26%)	639
Legalization of Abortion	105 (17%)	334 (55%)	164 (27%)	603

Table 1: Training data statistics. FAV = IN FAVOR, AG = AGAINST, and NONE = neither inference.

3 The Approach Overview

We decided to build a classical supervised learning system, in particular, we used a maximum entropy classifier (Loper et al., 2002). The classifier was trained separately for each topic. For the subtask B - weakly supervised system (*Donal Trump*) we used the *Hillary Clinton* training data, which we considered as the closest. We also added some enhancements we discuss later.

We first analysed hashtags in the training corpus. We automatically identified those that predict the stance well for each topic. These hashtags strongly correlate with one of the stance classes. For example, if the tweet contains hashtag *#benghazi* (the *Hillary Clinton* target), the stance is always AGAINST. We picked hashtags which appear at least in 10 tweets and at least 90% of these tweets are annotated with a particular stance.

The important hashtags are listed in Table 2.

We also gathered from Twitter additional data (via Twitter API) based on these automatically detected hashtags (#benghazi, #stophillary2016, etc.). This additional data was not used directly during the training phase but we created a set of dictionaries (ADSD) out of them.

Topic	Hashtag	Stance
Atheism	#freethinker	IN FAVOR
	#islam	AGAINST
Climate Change is a Real Concern	#climate	IN FAVOR
	#mission	IN FAVOR
	#peace	NONE
	#tip	IN FAVOR
Feminist Movement	#feminists	AGAINST
	#spankafeminist	AGAINST
Hillary Clinton	#benghazi	AGAINST
	#lol	AGAINST
	#stophillary2016	AGAINST
Legalization of Abortion	#alllivesmatter	AGAINST
	#ccot	AGAINST
	#prolifeyouth	AGAINST

Table 2: Hashtags analysis.

3.1 Preprocessing

Preprocessing starts the pipeline. Each of the following steps was applied to every tweet.

1. All URLs are replaced by keyword *URL*,
2. multiple exclamation marks are replaced by *MULTIPLEEXCLAMATIONS*,
3. mutiple question marks are replaced by *MULTIPLEQUESTIONMARKS*,
4. Twitter usernames like *@peter_krejzl* are replaced by *NAME*,
5. links to images (pic.twitter.com) are replaced by *IMGURL*,

6. hashtag *#sems* is removed,
7. initial tag *RT* is removed,
8. English stopwords are removed, [3]
9. only letters are preserved, the rest of the characters is removed,
10. for the *Donald Trump* target, we used training data from the *Hillary Clinton* target but we removed the following words from the tweets: *hillary, hilary, clinton*.

3.2 Features

A basic set of features was created from the preprocessed text. Unigrams perform quite well in the task (Somasundaran et al., 2009), so we used it as a baseline for all targets. The model is based on TF-IDF and uses not more than 750 features (first 750 words from the vocabulary). This is used for all five topics. Then, we implemented a set of other features that could be turned on or off for each topic.

We built a set of features from **hashtags** in Table 2. In maximum 50 unigram and bigram features were generated from the hashtags using the TF-IDF weighting.

Anand et al. (2011) showed that **initial n-grams** are useful features. Our system supports initial unigrams to initial trigrams, the maximum number is 50 features. However, from our experiments with the training dataset, we found useful only initial unigrams, and initial bigrams for the *Hillary Clinton* target (turned on for *Donald Trump* as well).

Another surface feature was **tweet length** (in words) after preprocessing.

Part-of-speech tags were generated from the preprocessed tweet and we built unigram and bigram data model using TF-IDF, limited to 50 features.[4]

General Inquirer (GI) [5] (General-Inquirer, 1966) provides dictionaries useful for example for sentiment analysis. We used a subset of the dictionary, in particular columns: *Positiv*, *Negativ*, *Hostile*, *Strong*, *Pleasure*, *Pain*.

Entity-centered sentiment dictionaries (ECSD): We used another resource borrowed from the sentiment analysis: dictionaries created mainly for the purpose of entity-related polarity detection (Steinberger et al., 2012). We used both the highly positive and positive terms[6] as IN FAVOR features and highly negative / negative terms as AGAINST features.

In some topics like *Legalization of abortion* or *Atheism* any **reference to a bible** (e.g., *Romans 12:2*) is also a very good indicator. We add additional binary feature based on the presence of a bible reference.[7]

Domain Stance Dictionary (DSD) Based on the training data analysis of each topic, we created a list of key words that tend to indicate a particular stance. We first generated a list of candidates: for each topic, we took words with ratio $frequency - in - topic/frequency - in - the - training - data > 0.6$ and $frequency - in - topic > 1$. If a word occurred at least 4 times more frequently in 'IN FAVOR' tweets than in 'AGAINST', it was added to the 'IN FAVOR' candidates' list. We repeated the same approach to produce 'AGAINST' candidates. The lists were then filtered manually and it resulted in strong stance-predictive keywords lists. All the lists together contain 221 words, an average list had 22 words. For instance, for the *Legalization of Abortion* topic, the following words or hashtags suggest the AGAINST stance: *unborn, womb, prolifegen, conception, precious, chooselife, kills, abortionismurder, destroys, itsnotonitsnotsafe, manslaughter, eliminated, cannibalism, heartbeat, ...* We used the number of words from each dictionary the tweets contain as features.

Additional Domain Stance Dictionary (ADSD): In the case of the *Legalization of Abortion* and *Hillary Clinton* topics, we created additional two dictionaries per topic. It was a similar exercise to DSD, but we used the additional tweets gathered through Twitter API as an input. For example, the AGAINST dictionary for *Hillary Clinton* contains: *attack, benghazihearings, blamed, blood, BloodOnHerHands, corrupt, irritated, Killary, ...*

[3] We used stopwords available in the *nltk.corpus* python library.

[4] We used *Nltk* part-of-speech tagger.

[5] http://www.wjh.harvard.edu/ inquirer/

[6] There are two levels of intensity for both polarities.

[7] Simple python regular expression `(\d+):(\d+)`.

Both the DSD and ADSD dictionaries contain terms that strongly indicate a particular stance. They were used to modify the final output towards the particular stance and override the classifier result. For each tweet we count the number of words from each positive or negative DSD/ADSD. If there are more words from the positive dictionary then the whole tweet is deemed FAVOR and vice versa. If the counts of positive and negative words are equal then the override logic is not used. We also noticed (during the development phase) that the classifier was overridden only few times in the Task A, while more times in the Task B. We think it is due to the different training data used for the Donald Trump task.

4 Configuration and Results

During the development phase, we used 10-fold cross-validation to test all combinations of features. Each particular dataset was split randomly. For each experiment we measured average F1-score on IN FAVOR and AGAINST classes, the same metric as the official one (Mohammad et al., 2016). This way we identified an optimal set of features for each topic, listed in Table 3.

Topic	**Features**
Atheism	Unigrams, Bible reference, DSD
Climate	Unigrams, Hashtags, POS, DSD
Feminism	Unigrams, Hashtags, DSD
Hillary	Unigrams, Hashtags, Initial unigrams, Initial Bigrams, ECSD, DSD, DSDA
Abortion	Unigrams, Bible reference, DSD, DSDA
Trump	Unigrams, Hashtags, Initial Unigrams, Initial Bigrams, ECSD, DSD, DSDA

Table 3: Features per topic used for the submission.

Table 4 shows results on the development set. We reached the best improvement over the baseline for *Hillary Clinton*, followed by *Feminism* and *Atheism*. However, detecting a correct stance for these targets seemed to be the most difficult.

Official results of the SemEval task are summarized in Table 5. There were 19 participating systems for subtask A and 9 for subtask B. We performed well for *Abortion* (2nd), *Climate* (3rd) and *Hillary Clinton* (4th) targets in comparison with the other participating systems, we received an average rank for *Atheism* and *Feminism*. The overall rank was 9th.

In the weakly supervised subtask (*Donald Trump*), we were ranked 4th, only the top system was significantly better. The difference between performances on the *Hillary Clinton* and *Donald Trump* topics (.5982 vs. .4202) indicate the difference in complexity between the subtasks.

It seems that although Clinton and Trump are competitors on political stage, Clinton's training data brought useful features for Trump as well (criticizing or praising a politician), although in many cases the overriding strategy corrected the classifier's prediction.

Topic	**Baseline (unigrams)**	**Features (no override)**	**Features + override**	**Diff against baseline**
Atheism	.5579	.6314	.6314	+13%
Climate	.6250	.6589	.6590	+5%
Feminism	.4722	.5424	.5443	+15%
Clinton	.4460	.5386	.5386	+21%
Abortion	.6250	.6749	.6749	+8%

Table 4: Development results per topic with features turned on/off (K-10 fold validation).

Topic	**UWB** (Rank)	**Avg. system**	**Best system**
Atheism	.5788 (8)	.5510	.6725
Climate	.4690 (3)	.4219	.5486
Feminism	.5182 (10)	.5155	.6209
Hillary	.5982 (4)	.5248	.6712
Abortion	.6198 (2)	.5472	.6332
Sub-task A	.6342 (9)	.6202	.6782
Sub-task B	.4202 (4)	.3737	.5628

Table 5: Official results of the SemEval task. There were 19 submissions for subtask A and 9 submissions for subtask B.

5 Conclusion

The paper describes our participation in the Tweets stance detection task of SemEval 2016. Our submission was based on a maximum entropy classifier with mainly surface-level, sentiment and domain-specific features. The system was among the top systems for three of the five targets from the supervised task. Without the labeled tweets, the weakly supervised scenario, our position was fourth (from nine). Currently, we investigate in more detail how to gather more training data automatically.

Acknowledgments

This work was supported by grant no. SGS-2013-029 Advanced computing and information systems and by project MediaGist, EU's FP7 People Programme (Marie Curie Actions), no. 630786.

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 WASSA'11*, ACL.

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., and Cherry, C. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of SemEval '16*, ACL.

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 Twenty-Seventh International Flairs Conference*, AAAI.

Schneider, J., Groza, T., Passant, A. 2014. A review of argumentation for the social semantic web. In *Semantic Web, 4(2)*, pages 159-218, IOS Press.

Hasan, K. S., and Ng, V. 2013. 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, ACL.

Kiritchenko, S., Zhu, X., and Mohammad, S. 2014. Sentiment Analysis of Short Informal Texts. In *Journal of Artificial Intelligence Research, vol. 50*, pages 723-762, AAAI Press.

Loper, E. and Bird, S. 2002. NLTK: The Natural Language Toolkit In *Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, ACL.

Mohammad, S., Kiritchenko, S. and Zhu, X. 2013. NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. In *Proceedings of SemEval-2013*, ACL.

Kiritchenko, S., Zhu, X. and Mohammad, S. 2016. Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text. In *Emotion Measurement*, Elsevier.

Mohammad, S., Kiritchenko, S. and Zhu, X. 2010. Support or Oppose? Classifying Positions in Online Debates from Reply Activities and Opinion Expressions. In *Proceedings of the International Conference on Computational Linguistics*, pages 869-875, ACL.

Rajadesingan, A. and Huan, L. 2014. Identifying Users with Opposing Opinions in Twitter Debates. Social Computing, Behavioral-Cultural Modeling and Prediction. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 153-160, Springer.

Recasens, M., Danescu-Niculescu-Mizil, C. and Jurafsky, D. 2013. Linguistic Models for Analyzing and Detecting Biased Language. In *Proceedings of ACL 2013*, pages 1650-1659, ACL.

Somasundaran, S. and Wiebe, J. 2009. Recognizing stances in online debates. In *Proceedings of the ACL/AFNLP*, pages 226-234, ACL.

Sridhar, D., Getoor, L. and Walker, M. 2014. Collective Stance Classification of Posts in Online Debate Forums. In *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media* pages 109-117, ACL.

Steinberger, J., Lenkova, P., Ebrahim, M., Ehrmann, M., Hurriyetoglu, A., Kabadjov, M., Steinberger, R., Tanev, H., Zavarella, V. and Vazquez, S. 2012. Creating Sentiment Dictionaries via Triangulation. In *Decision Support 53(4)*, pages 689-694, Elsevier.

Stone, P., Dumphy, D., Smith, M., Ogilvie, D. 1966. The general inquirer: a computer approach to content analysis. In *M.I.T. studies in comparative politics*, M.I.T. Press.

Thomas, M., Pang, B., and Lee, L. 2006. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of EMNLP*, pages 327-335, ACL.

Walker, M. A., Anand, P., Abbott, R., and Grant, R. 2012. Stance classification using dialogic properties of persuasion. In *Proceedings of the NAACL/HLT*, pages 592-596, ACL.

Wyner, A., and Schneider, J.. 2012. Arguing from a Point of View. In *Proceedings of the First International Conference on Agreement Technologies*, CEUR.