ltl.uni-due at SemEval-2016 Task 6: Stance Detection in Social Media Using Stacked Classifiers

Michael Wojatzki
Language Technology Lab
University of Duisburg-Essen
Duisburg, Germany
michael.wojatzki@uni-due.de

Torsten Zesch
Language Technology Lab
University of Duisburg-Essen
Duisburg, Germany
torsten.zesch@uni-due.de

Abstract

In this paper, we describe our participation in the first shared task on automated stance detection (*SemEval 2016 Task 6*). We consider the task as a multidimensional classification problem and thus use a sequence of stacked classifiers. For subtask A, we utilize a rich feature set that does not rely on external information such as additional tweets or knowledge bases. For subtask B, we rely on the similarity of tweets in this task with tweets from subtask A in order to transfer the models learnt in subtask A.

1 Introduction

Stance-taking is an essential and frequently observed part of online debates and other related forms of social media interaction (Somasundaran and Wiebe, 2009; Anand et al., 2011). In the *SemEval 2016 Task 6: Detecting Stance in Tweets* (Mohammad et al., 2016), stance is defined relative to a given target like a politician or a controversial topic. A text can then either be in favor of the given target (FAVOR), or against it (AGAINST). As the dataset also contains texts without a stance, we additionally have to deal with the the class NONE.

Being able to automatically detect and classify stance in social media is important for a deeper understanding of debates and would thus be a great tool for information seekers such as researchers, journalists, customers, users, companies, or governments. In addition, such analysis could help to create summaries, develop a deeper understanding of online debating behavior, identify social or political groups, or even adjust recommendations to users' standpoints (Anand et al., 2011; Sridhar et al., 2014; Boltuzic and Šnajder, 2014).

In the following, we describe our system for stance detection. We did not make use of any sources of external information such as additional tweets or stance knowledge bases, as our goal was to rely only on the provided training data. Since the task allowed just for one submission, we include some further analysis that will shed light on the usefulness and impact of the used features and parameters.

2 Subtask A – Supervised Framework

The goal of this subtask is to classify tweets about five targets: *Atheism*, *Climate Change is a Real Concern*, *Feminist Movement*, *Hillary Clinton*, and *Legalization of Abortion*. For each target, there are about 400-600 manually labeled tweets that can be used for training.

As the targets are quite different, we train a separate classifier for each of them. Additionally, we split the three-way classification into a stacked classification, in which we first classify whether the tweet contains any stance (classes FAVOR and AGAINST) or no stance at all (class NONE). In a second step, we classify the tweets labeled as containing a stance as FAVOR or AGAINST. This sequence of classifications is visualized in Figure 1.

All shown classifications are implemented using the DKPro TC framework[1] (Daxenberger et al., 2014) and utilize the integrated Weka SVM classifier.

[1] version 0.8.0-SNAPSHOT

Proceedings of SemEval-2016, pages 428–433,
San Diego, California, June 16-17, 2016. ©2016 Association for Computational Linguistics

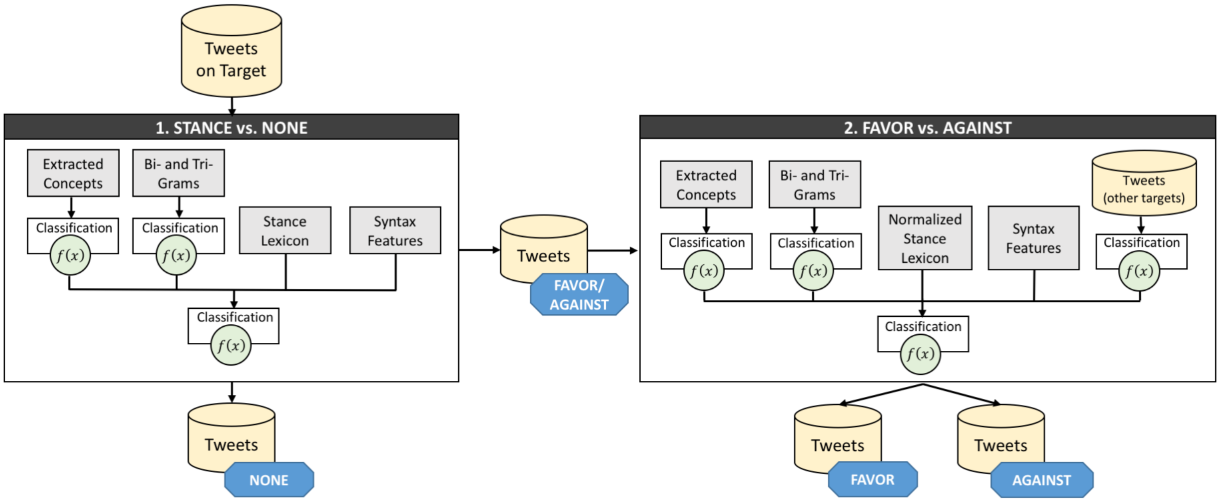

Figure 1: Overview on the sequence of stacked classifications that is used for the supervised setting (subtask A)

2.1 Preprocessing

We use the DKPro Core framework[2] (Eckart de Castilho and Gurevych, 2014) for preprocessing. We apply the twitter-specific tokenizer Twokenizer[3] (Gimpel et al., 2011), the DKPro default sentence splitter, and the Arktweet Pos tagger[4] (Gimpel et al., 2011).

As the Arktweet PoS tagger has a special tag for hashtags, even syntactically integrated hashtags like in *I like #Hillary !* would be assigned the tag *hashtag*. Since our feature set relies on the syntactic role of the used words (e.g. nouns as the subject or object of an proposition), we only keep hashtag labels that occur at the end of a tweet. For all other hashtags, we additionally apply the OpenNlp PoS tagger[5] and overwrite the hashtag label with the syntactic category.

Afterwards, we annotate a fixed set of negations (*not*, *no*, *none*, *nor*, *never*, *nobody*, *neither*, *nowhere*) including contractions such as *can't*, *aren't*. Finally, we annotate modal verbs (*can*, *could*, *may*, *might*, *must*, *shall*, *should*, *will*, *would*).

2.2 Features

Figure 1 shows that both classifications use roughly the same feature set. However, the FAVOR vs. AGAINST classification additionally uses a feature that transfers models learnt from other targets. In the following, we describe all features in detail, explain how they may relate to stance taking language, and outline differences in both classifications.

N-Gram Features We use the 500 most frequent *bi*-and *trigrams* as binary features to capture expressions that are longer than single words. Thereby we want to approximate multi word expressions such as *climate change*. As the resulting 500 features would outnumber the other features, we use stacking to classify according to the n-gram features only and use the outcome as a single feature in the overall model. Note that we handle *unigrams* by using an automatically created stance lexicon.

Syntactic Features According to Faulkner (2014), the usage of conditional sentences and modal verbs may indicate stance taking behavior. Hence, we use the number of sentences starting with *if* and the occurrence of modal verbs as a feature. As stance-taking behavior may be indicated by the usage of exclamation- and question marks (Anand et al., 2011), we use as a feature the overall counts as well as the count of over-usage like *???* or *!?!*. Finally, we use as a feature the number of negation markers in a tweet.

Stance-lexicon Features For each target, we create a unigram stance lexicon by computing their statistical association with one of the outcome val-

[2] version 1.7.0
[3] version 0.3.2
[4] version 0.3.2
[5] *maxent* model in version 3.0.3

ues. This feature is inspired by the work of Somasundaran and Wiebe (2010) who use a subjectivity lexicon that was created by using the statistical association with negative and positive argumentation. We compute the association measure $gmean$ (Evert, 2004) of every unigram towards the two poles (FAVOR/AGAINST or STANCE/NONE) and then use the difference between both values as the stance score s. The $gmean$ association of a word x with a polarity class $+$ is computed as:

$$gmean_{+}(x) = \frac{c_{+}(x)}{\sqrt{c_{+}(x) \cdot c_{-}(x)}} \qquad (1)$$

where $c_{+}(x)$ is the count of x in $+$ and $c_{-}(x)$ is the count of x in $-$. Based on the computed stance lexicon, we calculate the final polarity of as the normalized sum over the stance score s for each token in a tweet. In contrast to the mere use of *unigram* features the resulting lexicon has the advantage that it can distinguish between words that indicate stance with varying strengths.

As a defining feature of social media is the frequent usage of nonstandard spelling, we use text normalization in order to make this feature more robust. We first lowercase all tokens, remove @- and #-prefixes, and lemmatize plurals. For tokens that still do not appear in our stance lexicon, we compute the normalized *Levensthein* edit distance (implemented using the DKPro Similarity (Bär et al., 2013)) and use the largest score if the value is smaller than .15. However, the normalization is only applied for the STANCE vs. NONE classification, since we observed that even capitalization can be a signal for being in favor or against a certain target.

In addition to unigram stance lexicons, we create a hashtag stance lexicon by using the same methodology, but just considering the tokens with the hashtag PoS tag (see Section 2.1).

Concept Features When analyzing the data, we observed that each target is associated with a few concepts that are subject of a controversial debate. Whereas the stance lexicon models words that are highly associated with one class, these words are associated with both classes. Since they are used by authors of different stances, they can be be considered as being central for the debate. In order to retrieve the most central concepts, we select the top 12 nouns from each target. These candidates are then normalized (in the same fashion as described above) and cleaned for concepts that are not controversial (i.e. they are only used by authors with the same stance). For reasons of automation, no manual revision of concepts (e.g. handling words that are parts of multiword expressions such as *climate* and *change* or terms that are semantically related such as *feminist* and *feminism*) had been done. The remaining concepts are shown in Table 1.

For these concepts we learn, whether they express being in favor or against (respectively a stance at all) depending on their context. In order to model the context, the classifiers are equipped with n-gram features (*top 200 uni-*, *bi-*, *trigrams*). Finally, the prediction of these concept-classifiers is used as a feature.

Target Transfer Features Some targets appear to have a certain thematic overlap (e.g. there seems to be an overlap between *Legalization of Abortion* and *Feminist Movement* because both concern the rights of women). For example, consider a tweet about the target *Feminist Movement* that contains a stance towards abortion. If we want to classify the stance towards *Feminist Movement* it seems naturally to incorporate the stance towards abortion. This idea is modeled by applying models to a target that have been trained on a different target. The resulting classification is then used as a feature.

We only apply this feature if the tweet has a minimal topical overlap with one of the other targets domains. We found that on the training data, it works best if a tweet is related to a target if it contains one of the top 60 frequent nouns or named entities for some target domain (see also Section 3).

As shown in figure 1, this feature is only used for the AGAINST vs. FAVOR classification. On the training data, this feature only had an impact for the targets *Climate Change is a Real Concern*, *Hillary Clinton* and *Legalization of Abortion*. Thus, we only apply it for these targets on the test data, too.

2.3 Results

We report gained results on the provided test data using the official metric that is the macro-average of F_1(FAVOR) and F_1(AGAINST). As shown in the first row of Table 2, our system achieved a score of

Atheism	Climate Change	Feminist Movement	Hillary Clinton	Abortion
day	change	equality	campaign	abortion
death	climate	feminism	candidate	baby
faith	summer	feminist	country	body
god		gamergate	hillary	child
life		gender	hillaryclinton	choice
religion		male	president	life
		man	support	time
		rape	time	woman
		time	vote	
		woman	woman	

Table 1: Extracted concepts for which a separate classifier is trained

.62. This corresponds to rank 12 in the official ranking (Mohammad et al., 2016). The performance of our system varies significantly between the different targets. For instance, the difference between the classification of the target *Legalization of Abortion* and *Climate Change is a Real Concern* is about 20 percent points.

In order to analyze the impact of the used features, we conducted an ablation test. The results in Table 2 show that the stance lexicon is the only feature that has a significant impact on all targets. The concept features seem to be helpful for some targets. When training a model with only those two features, we reach a score of .65 (whole test data) compared to .62 when using all features.

3 Subtask B – Weakly Supervised Framework

There is no training data for subtask B, so this has to be tackled in an unsupervised or weakly supervised fashion. The target is *Donald Trump* and participants are provided with large corpus (about 78 000 tweets) of un-annotated tweets.

Our approach for task B works in two stages (in analogy to task A): First, in order to determine whether a tweet has any stance at all, we compare each tweet with the whole collection. We found on subtask A, that if one filters all tweets that do not contain one of the n most frequent nouns or named entities of the target, the majority of the remaining tweets have a stance. Of course, larger values for n will improve recall on the cost of precision. As the second classification (FAVOR vs. AGAINST) depends on a high precision, we decided to use a threshold that assigns a higher weight to precision. We empirically found that $n = 60$ works well on subtask A. Tweets that are not similar are treated as NONE from here on.

Second, we select all targets (from Subtask A) that are most similar to a tweet. If a tweet is not similar to any target, we considered its stance as UNKNOWN from here on. If a target is selected, we use the model that has been trained on it to classify the tweet. This is the weakly-supervised part, as we apply a model that is trained on a different target.

There is one additional complication: One may argue that there is an inverse relationship between the stances of subtask A and the target *Donald Trump*. For instance, if a tweet is in favor of *Hillary Clinton*, there is a high likelihood that the tweet is implicitly against *Donald Trump*. A similar relation may be assumed for the other targets or at least for the majority of his supporters. Consequently, all classifications are inverted (AGAINST becomes FAVOR and FAVOR becomes AGAINST) and summed up. The final decision is then made on the majority vote. In case there are as much votes for FAVOR as for AGAINST we assume that the stance is UNKNOWN.

Results Our approach yields a score of .26 which corresponds to the 11^{th} rank. However, this is inferior to the the performance of the base-rate (AGAINST) which is about .3. This suggests that the made assumptions (e.g. inverse relationship between stances of subtask A and B, majority vote of classifiers) are not optimal.

	All	Atheism	Climate Change	Feminist Movement	Hillary Clinton	Abortion
all features (SemEval Submission)	.62	.53	.36	.55	.44	.57
- stance lexicon	.54	.48	.29	.50	.41	.46
- concepts	.62	.52	.35	.54	.46	.58
- negation	.62	.55	.36	.55	.44	.57
- target transfer	.62	.53	.35	.55	.46	.58
- punctuation	.62	.56	.35	.55	.47	.57
- conditional sentences	.63	.58	.36	.55	.44	.60
- modal verbs	.63	.58	.36	.55	.44	.60
- n-grams	.64	.59	.38	.56	.51	.58

Table 2: Ablation test of the feature set on the test data

4 Conclusion

In this paper, we presented our approach on automated stance detection based on stacked classifications. We split the three-way classification into a first classifier deciding whether there is any stance at all, and a second classifier that only makes a decision about the polarity of a tweet with a stance. Overall, we found a significant variation in performance across the targets and even between train and test data. An ablation test shows that from our rich feature set, only the automatically derived *stance lexicon* feature has a significant impact.

In general, our system (as well as all other participating systems) leaves much room for improvement. Stance detection could probably benefit from more semantically oriented features that go beyond the surface form of words.

Acknowledgments

We would like to thank the the organizers of the SemEval task 6 for this interesting task and the three annonymous reviewers for their helpful comments. This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under grant No. GRK 2167, Research Training Group ”User-Centred Social Media”.

References

Pranav Anand, Marilyn Walker, Rob Abbott, Jean E Fox Tree, Robeson Bowmani, and Michael Minor. 2011. Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis*, pages 1–9. Association for Computational Linguistics.

Daniel Bär, Torsten Zesch, and Iryna Gurevych. 2013. Dkpro similarity: An open source framework for text similarity. In *ACL (Conference System Demonstrations)*, pages 121–126.

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

Johannes Daxenberger, Oliver Ferschke, Iryna Gurevych, and Torsten Zesch. 2014. Dkpro tc: A java-based framework for supervised learning experiments on textual data. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 61–66, Baltimore, Maryland, June. Association for Computational Linguistics.

Richard Eckart de Castilho and Iryna Gurevych. 2014. A broad-coverage collection of portable nlp components for building shareable analysis pipelines. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT*, pages 1–11, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Stefan Evert. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, IMS, University of Stuttgart.

Adam Faulkner. 2014. Automated classification of stance in student essays: An approach using stance target information and the wikipedia link-based measure. In *Proceedings of the 27th International Florida Artificial Intelligence Research Society Conference*, pages 174–179.

Kevin Gimpel, Nathan Schneider, Brendan O’Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Pro-*

ceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2, pages 42–47. Association for Computational Linguistics.

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, June.

Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 226–234. Association for Computational Linguistics.

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124. Association for Computational Linguistics.

Dhanya Sridhar, Lise Getoor, and Marilyn Walker. 2014. Collective stance classification of posts in online debate forums. *ACL Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, page 109.