Status Report

Michael Wojatzki

Abstract

This document gives an overview on the progress of my PhD covers the following aspects: First, I briefly outline theoretical background and important assumptions of the project. Thereby, I will discuss current state of our knowledge and do not reflect how it has developed over time. Second, I describe a methodological framework in which we can arrange the conducted research. While this framework does not completely fit all of the conducted research, I try to be as comprehensive as possible. Using this framework, I subsequently reflect on results which are obtained in the so far executed experiments. As requested by the last reviews, I will thereby clearly differentiate between published, unpublished but finished and unfinished works. Finally, this report closes with remarks regarding future work and on professional activities such as networking and organized workshops.

1 Introduction

In my PhD project I am dealing with the automatic analysis of discussions and debates conducted on social media platforms. Besides developing and improving algorithms that are capable of detecting arguments, positions, substantiations, etc., goal of the project is to utilize these algorithms to contribute to a better understanding of social media discussions. Being capable of automatically analyzing social media discussions aids a huge variety of practical applications 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; Boltužić and Šnajder, 2014).

My work is located at the interface between different strands from the research field of natural language processing or computational linguistics. The first strand is called is called *Argument Mining* and is concerned with the identification of argumentative structures within written communication (Green et al., 2014). Approaches of argument mining typically assume highly complex and specialised argumentative structures (such as Toulmin (1958) or Freeman (1991)) and are merely applied to well structured text forms such as legal documents, scientific writing, etc. However, text harvested from social media is usually shorter, less dense for arguments, noisy and the used arguments are often not as sophisticated.

The other stand is called *opinion mining* or *sentiment analysis* which tries to identify the polarity of a given text. Sentiment analysis can be further distinguished (e.g. by Pang and Lee (2008) or Liu (2012)) into a) document-level or sentence-level analysis which wants to determine whether a text has a positive or negative sentiment (Pang et al., 2002; Turney, 2002; Yu and Hatzivassiloglou, 2003) and b) aspect-based sentiment analysis (Hu and Liu, 2004). The goal of aspect-based sentiment analysis has been formulated as identifying the aspects (e.g. price or quality) of given target entities (e.g. laptops or restaurant) and the polarity expressed towards them (positive, negative or neutral) (Pontiki et al., 2014; Pontiki et al., 2015; Pontiki et al., 2016). A variety of approaches have demonstrated that sentiment analysis is applicable to social media and its specific characteristics. For instance, since 2013 there is track on sentiment analysis in Twitter which annually receives a large number of submissions. However, sentiment analysis is usually limited to direct linguistic expressions of liking or disliking an entity its aspects (Pontiki et al., 2014; Pontiki et al., 2015; Pontiki et al., 2016) and thus often misses important parts of discussions such as allusions or conclusions.

In addition, after analysing the state-of-the-art in both strands, we identified implicitness as an crucial aspect that has been widely ignored so far. Implicitness is problematic as, in contrast to elaborated text, social media contains a high proportion of implicit discussions (e.g. see Habernal et al. (2014) or Boltužić and Šnajder (2014)) which often rely on community specific, shared assumptions. While models of sentiment analysis typically ignore implicit expressions or inferences (e.g. see (Pontiki et al., 2014; Pontiki et al., 2015; Pontiki et al., 2016)), models in argument mining are very sensitive to missing (but implicitly present) text parts (Habernal et al., 2014).

A related model that incorporates implicitly is the model of *stance*. The task of stance detection has been identified as the automated determination of whether an author expresses to be in favor (⊕) or against (⊖) a certain target, or if none of these conclusions is reasonable (NONE) (Mohammad et al., 2016a). Thereby, the *target* of the debate can be an entity (e.g. a politician) or any other controversial issue such as the *death penalty, climate change, etc.*. However, similar to sentiment analysis, *stance* is defined only at the document or utterance level, and therefore does not offer the possibility to model differentiated standpoints (e.g. expressing both points in favor and against the debate target in one utterance, giving justifications, etc.). In order to overcome these shortcoming, we propose a new model to capture debates on social media platforms. In the following, I will describe this model and how it is connected to related work.

2 Stance Based Argument Mining

In order to overcome the challenges which are described above, we introduce a new model based on a *debate stance* that will in most cases be implicit, but can be inferred from one or more *explicit stances* that rely on textual evidence from the utterance. We thereby assume that an utterance is always made in the context of a certain debate. Figure 1 exemplifies the assumed model and also shows it's advantages compared to a claim-premise model. In implicit argumentation the claim usually needs to be inferred, as it is not explicitly expressed (see figure 1b). We argue, that in the absence of explicit information, the claim always corresponds to the overall stance in the debate in which the utterance is made.

In the context of a debate about atheism, an utterance like *God will judge those infidels!* expresses a stance in favor of a supernatural power (*Supernatural Power* ⊕), while the actual stance on the debate target of atheism (*Atheism* ⊖) is not visible but must be inferred. Note that the debate stance might also be explicitly expressed (see figure 1a), but in implicit argumentation it has to be derived from the explicit stances.

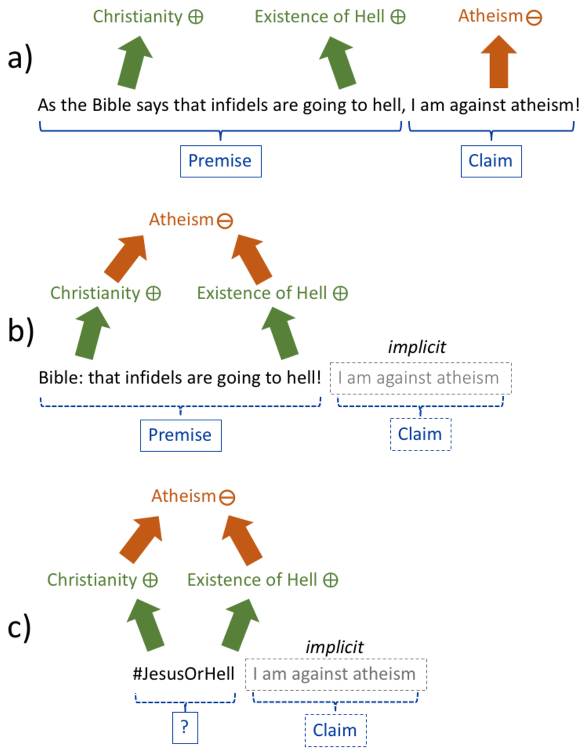

Figure 1: Stance-based vs. *claim-premise* model

In principle, each utterance evokes a large set of implicit stances (in a similar way as the iceberg contains a lot of invisible ice below the waterline). For instance, one may infer that a person uttering *Bible: infidels are going to hell!* is probably in favor of praying and might have a negative stance towards issues such as abortion, same-sex marriage, etc. However, we argue that being in favor of Christianity already implicitly covers these stances under a common sense interpretation. Depending on the present informational need these targets may be more or less relevant.

Thereby, our model of implicit argumentation aligns with the *Relevance Theory* proposed by Sperber and Wilson (1986) and the *Cooperative Principle* by Grice (1970) as we also assume that utterances provide hints on the intended meaning to the recipient. Particularly, our model shares the assumption of *Relevance Theory* that the precision of statements is such that a receiver can decode the meaning only by incorporating the context.

2.1 Related Work

In this related work section, I will briefly point to approaches and studies that serve as foundations of the proposed model and my project.

2.1.1 Differentiating Stance

In order to differentiate stance in debates, one often relies on on sub-structures, which are more explicit than the debate stance. Controversial substructures in debates and their relations to stance have been studied by several researchers. We distinguish these works based on the origin of the sub-structures. Swanson et al. (2015) and Misra et al. (2015) extract sub-structures using text summarization techniques and then group them by a similarity measure which incorporates stance. For instance, if two statements relate to the same target, but express different polarities they are considered to be *roughly equivalent*. Consequently, stance is modelled only indirectly but may be inferred from the grouping of statements by the similarity measure. In addition, their approach relies on text summarization which does not make sense for very short texts such as the shown examples. Hasan and Ng (2014) use reoccurring sub-structures on which several annotators have agreed. Conrad et al. (2012) manually model two hierarchies of argumentative phrases with positive and negative stance as root nodes. Each hierarchy consists of more general phrases (such as *bill is politically motivated*) which are refined by phrases in the lower level of the hierarchy. Sobhani et al. (2015) analytically define sub-debates, while Boltužić and Šnajder (2014) extract them from the debating website *idebate.org*.

2.1.2 Stance Detection

Besides having an expressive and reliable model, it is important to keep in mind that models also have to prove their practicality. However, as approaches that fully automatize a model with both debate stance and explicit sub-structures are rare and the corresponding evaluations are highly tuned towards the used dat set and annotations. Since there are plenty of comparable approaches for the underlying technique of automated stance detection, I will focus on describing the state-of-the-art in this area.

Most state-of-the-art approaches for stance detection are based on supervised machine learning. Supervised learning refers to automatically inferring a classification function based on a set of training examples.

Early approaches in stance detection adapt supervised sentiment classification techniques to distinguish between $\ominus$ and $\oplus$ stance in texts. They rely on classifiers equipped with engineered features such as ngrams, negators, modal verbs, lexicons or dependency features (Walker et al., 2012; Anand et al., 2011; Hasan and Ng, 2013; Faulkner, 2014). As they do not consider the NONE class, we argue that they are not suitable for real-world applications, as these are usually confronted with noisy data and need to filter out irrelevant texts. The consideration of the NONE class was recently established by the shared tasks on stance detection SemEval 2016 task 6 (Mohammad et al., 2016a) or the NLPCC Task 4 (Xu et al., 2016).

While the participants in the shared tasks have used a variety of approaches, we identify two main strands. First, knowledge-light, neural network architectures as used by the best teams in the SemEval challenge (Zarrella and Marsh, 2016; Wei et al., 2016). These approaches translate the training data in sequences of pre-trained word embeddings and feed them into neural networks. More precisely, the top scoring approach utilizes a long-short-term-memory (LSTM) layer (Zarrella and Marsh, 2016). LSTMs are an augmentation of recurrent neural networks which utilize a special gate node (called forget gates). This gate node allows the LSTM layer to abstract features based on sequences with varying length. Second, more traditional classifiers which represent the data mostly through word and character ngrams, averaged word-embeddings and sentiment features. These representations are than used to train models with algorithms such as support vector machines (SVMs) (Mohammad et al., 2016a; Xu et al., 2016). While the results of SemEval and NLPCC show that these approaches outperform the neural approaches, neural approaches are highly competitive which underlines that it is ultimately unclear which strand is superior.

3 Methodology/Data

In order to ensure, that the made assumptions are reasonable, it is important examine the alignment to actual, real-world data. In doing so, I attach great importance to covering a variety of different data sets to cover variability with regard to targets and domains. By comparing methods on different data, we try to derive more general findings.

Since only a few (resp. no) data sets are available which meet our requirements, we also construct data ourselves by conducting annotation experiments. Annotation studies typically include $>= 2$ annotators who try to apply a annotation scheme (a model) to the same data. By comparing the annotators, these studies allow valuable insights into the reliability of the annotation, which can additionally provide evidence for the validity of our model.

Annotated data also enables an examination of quantitative dependencies between the components of our model. These patterns correspond to the interpretability of our model, as they enable to develop a deeper understanding of social media discussions. These patterns include the frequency to which certain explicit stances occur, several explicit stances co-occur and to which explicit stances occur with a certain debate stance.

Finally, using the annotated data, we can develop and test algorithms, which is a major goal of this work.

We currently have worked with the following data sets:

SemEval 2016 - Task 6 (Mohammad et al., 2016b) This data set is part of the first shared task on automated stance detection. The organizers provided a trial, training, and test dataset of 100, 2814, and 1249 tweets, respectively for detecting stance towards five targets namely *Atheism*, *Climate Change is a Real Concern*, *Feminist Movement*, *Hillary Clinton*, and *Legalization of Abortion*. In addition, the organizers provide an unlabelled corpus and labeled test-data towards the target *Donald Trump*.

Atheism Stance Corpus (Wojatzki and Zesch, 2016b) In this study, we reannotate tweets on *Atheism* (513 from the training set and 220 from the test set), as we found the topic to require less knowledge about specific political events from the SemEval task. We let three annotators (two undergraduate and one graduate student of cognitive science) identify stances towards the explicit targets and the debate target.

To derive explicit targets we utilize a semi-automated, bottom-up approach that focusses on concepts that are mostly explicitly expressed by named entities and nouns. As we want to ensure that the targets differentiate the authors' positions sufficiently, we consider the degree of association between concepts and stance polarities. In detail, we compute the collocation coefficient $Dice$ (Smadja et al., 1996) for each word, and selected the 25 words which are most strongly associated with *Atheism* $\ominus$ and *Atheism* $\oplus$. As the resulting concepts are too numerous and too fine-grained to be used in our model and ,thus, manually them into more coarse-grained targets.

YouTube Stance Corpus (Wojatzki and Zesch, 2017) This newly created data set contains youtube comments towards death penalty. First, we apply a semi-automated selection to find videos that are as representative as possible and obtained a set of six videos. Then we polled the Youtube API[1] and received 821 comments (313 of them replies) from 614 different users with a total of 30 828 tokens. Using the same annotation scheme as in the SemEval task, we had three graduate students annotate each comment with stance towards death penalty and explicit stances. As explicit targets, we rely on targets from the debating website *idebate.org*[2] and from social media debates on *reddit.com*. For the reddit targets, we extracted targets from the subreddit [3], where users post controversial standpoints and invite others to challenge it.

Hatespeech Corpus (Ross et al., 2016) As I argue that one can consider hate speech as an extreme, unpleasant form of stance taking, I also report on the resources that were created in the interdisciplinary working group *hate speech*. In the IWG we have so far conducted two studies and corresponding data sets. In the first study (Ross et al., 2016), we created a german hate speech corpus and conducted an experiment in which we demonstrate that hate speech is a (too) fuzzy concept and that is very hard to reach consensus on whether a tweet is hate speech or not. In the second study (Benikova et al., 2017), we shed light on the influence of implicitness on hate speech annotations and show an measurable influence which is however, moderated by e.g. content variables.

[1] `http://developers.google.com/youtube/;` v3-rev177-1.22.0

[2] `http://idebate.org/debatabase/debates/capital-punishment/house-supports-death-penalty`

[3] `http://www.reddit.com/r/changemyview`

4 Published Results

4.1 ltl.uni-due@SemEval Task 6 (Wojatzki and Zesch, 2016a)

We participated in the *SemEval 2016 Task 6: Detecting Stance in Tweets* which represents the first shared task on stance detection that tried to explore the state-of-the-art.

Figure 2 gives a overview on our system for automated stance detection (for a a more detailed explanation see Wojatzki and Zesch (2016a)).

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 $\oplus$ and $\ominus$) or no stance at all (class NONE). In a second step, we classify the tweets labeled as containing a stance as $\oplus$ or $\ominus$. Our analysis revealed that simple word features are best suited to learn the relationship between tweets and stance.

As demonstrated by the task organizers none of them beats the provided baseline which is a support vector machine equipped with character and word ngrams (Mohammad et al., 2016a). In a post-hoc analysis they show that this baseline profits from leveraging unlabelled data (Mohammad et al., 2016b). In general, the results suggest that the performance of the state-of-the-art is significantly better than a random or a majority class baseline, but still leaves huge room for improvements.

4.2 Stance-based Argument Mining (Wojatzki and Zesch, 2016b)

In this publications we postulated the model which is explained in section 2. As explained in the previous section, we applied this model to a sub set of the SemEval data. Figure 3 shows the Fleiss' κ for the annotation.

As shown in figure 3, we obtain a Fleiss' κ of 0.72 for the annotation of the debate stance which is comparable to scores in the literature. Two explicit targets (*Christianity* and *Islam*) yield especially high agreement, because they are associated with clear signal words such as *Jesus* and *Quran*. Other targets such as *Secularism* are rather abstract. They hardly involve special signal words but still gain high agreements, which shows that our annotators did not just recognize certain keywords, but also reliably annotate more abstract targets. An error analysis for the target *Same-Sex Marriage* shows that there is disagreement if the tweet contains a stance towards gay rights in general but not to gay marriage. Finally, we obtain a κ of 0.63 for the joint decision on both the debate and the explicit targets.

Patterns In order to inspect usage patterns of explicit stance taking, we must agree on one annotation for each tweet. Since we do not assume that there are differences in the quality of the three annotators, we rely on a majority vote to compile a final annotation.

Figure 4 visualizes the frequency of the explicitly taken stances for Atheism $\oplus$ and Atheism $\ominus$. It shows that there are significant differences in the argumentation patterns between the two camps. As expected, if advocates of atheism are against a target such as *Christianity*, the opponents are mostly in favor of it or do not mention it.

From these analysis we can conclude that stable patterns of argumentation using explicit stances other than the debate stance exist. This is a strong indication for the validity of our assumption that the debate stance can be inferred from explicitly expressed stances.

Machine Learning In order to investigate how well our model can be assigned automatically, we conduct classification experiments and compare with suitable baselines. Based on how well the components are classifiable, we can derive how well the model is assignable as a whole.

We re-implement a state-of-the-art classifier as described in section **??**, run experiments in the manner of a ten-fold cross-validation and report micro averaged F_1. Besides the majority class baseline ($F_1 = .49$), we use the same setup as for the explicit stances to train an n-gram based classifier and obtain an F_1 of .66. In order to evaluate the usefulness of explicit stances for inferring the debate stance, we use the predictions from the previous experiment as features. This stacked classifier performs on par (.65) with the n-gram based classifier. It seems that the quality of predicting explicit stances is not yet good enough to improve over the state-of-the-art without incorporating general n-gram features. To estimate the potential of explicit stance features for classifying the debate stance, we add an oracle condition to our experiments in which we assume that the classification of explicit stances is done correctly. This classifier using only the manually annotated ex-

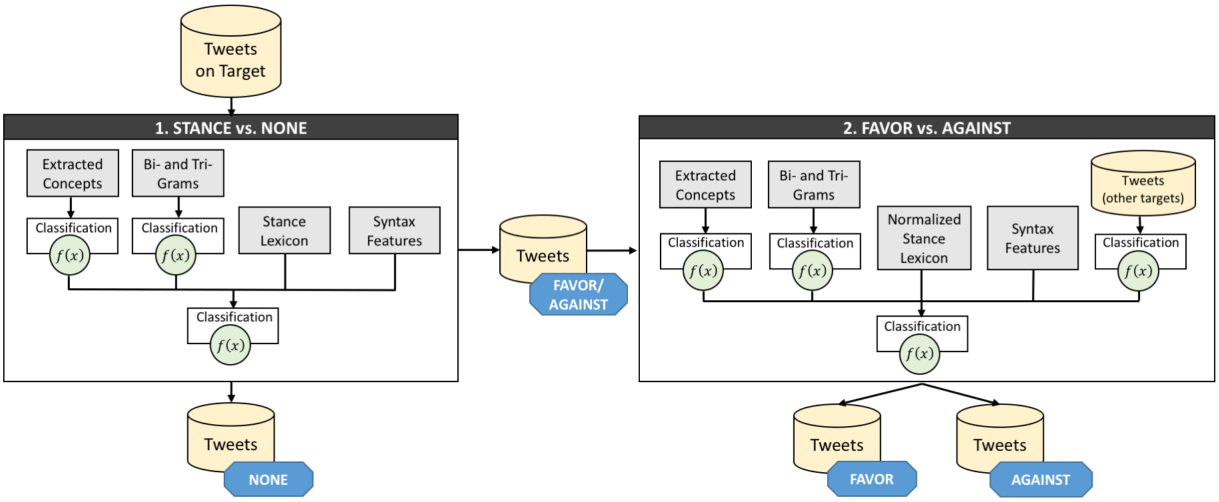

Figure 2: Overview on the sequence of stacked classifications that is used for the supervised setting (subtask A)

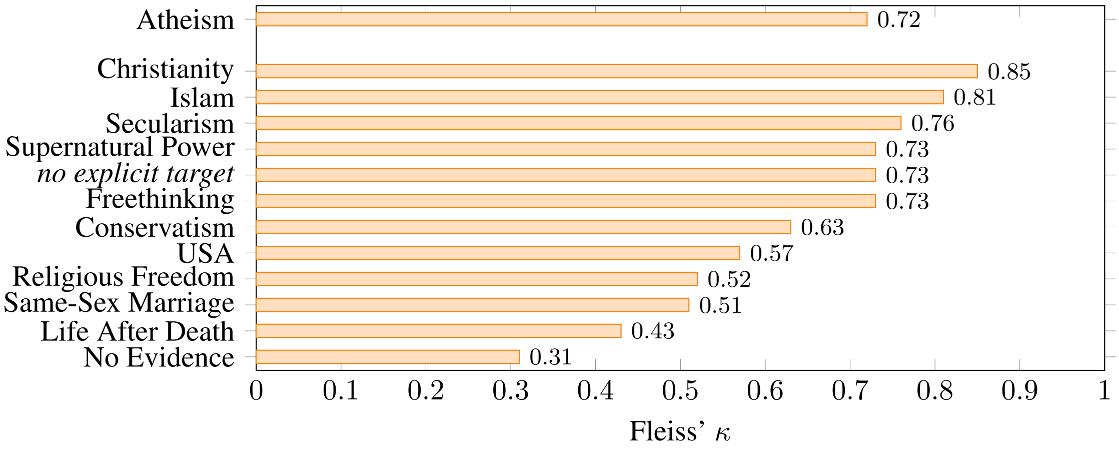

Figure 3: Inter-annotator agreement of the debate stance *Atheism* and explicit stances

plicit stances yields an F_1 score of .88 showing that large improvements over the state of the art are possible if explicit stances can be more reliably classified. We believe that this is indeed possible as explicit stances are always grounded in the text itself, while the debate stance might only be indirectly inferred.

5 Unpublished Results/ Under Review

5.1 Neural and Non-neural Stance Classifiers learn Sub-debates: A Study on Death Penalty Comments on YouTube (Wojatzki and Zesch, 2017)

For our annotation on the YouTube data set, we also compute *Fleiss' κ* between the three annotators to assess reliability. These scores are shown in Figure 5. For the debate stance, we obtain a value of .66 which is in a similar range as the comparable annotations of Sobhani et al. (2015) who report a weighted κ of .62 and Wojatzki and Zesch (2016b) who report a *Fleiss' κ* of .72.

Overall, we obtain a mixed picture for the annotation of the sub-debates, as we get κ values in a range from .13 up to .87. While the majority of sub-debates is annotated with a κ of above .6, there are significant deviations downwards such as *Financial Burden* with a κ of .26. With respect to the sets, there are few differences, as both contain sub-debates with low and high reliability.

Patterns As in the Atheism corpus, we again inspect quantitative patterns. These dependencies are visualized in Figure 1. Slightly over 50% of comments do not have any explicit target, but this is mainly due to the NONE cases that do not ex-

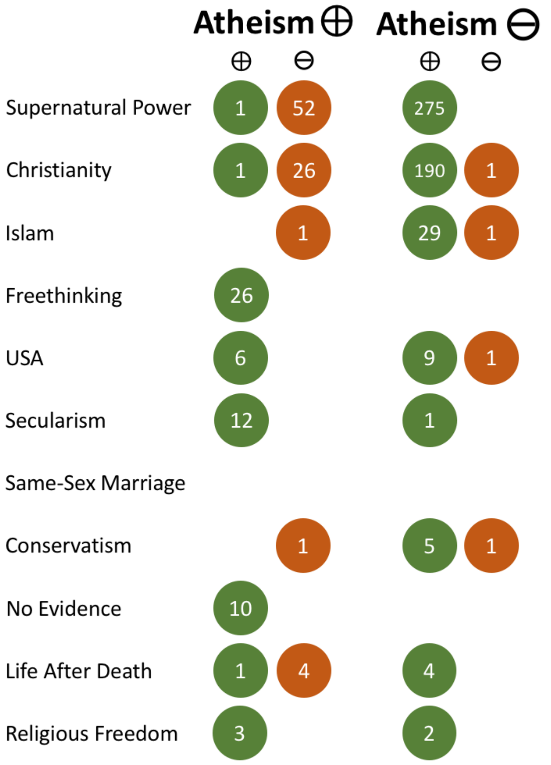

Figure 4: Frequency of explicit stances grouped according to debate stance

press any stance at all. For the polar instances (⊖ + ⊕), 389 out of 496 comments (78%) have one or more explicit target.

In the main diagonal of the table, we visualize how often a certain explicit target occurrs in the data. Overall, none of the two inventories can be clearly preferred over the other, as both contain frequent and thus useful sub-debates and ones that occur infrequently or not at all.

We observe, that the most frequent sub-debates (*Heinous Crimes* and *Eye for an Eye*) are often used as proxies for a general position in the debate, such as in the comment *If someone raped and killed your family, do you think they should continue to exist?* The second most common group are more specific sub-debates (e.g. *Deterring*) which occur in about 10% of the polar instances. The other sub-debates seem to cover rather marginal positions in the debate.

Another interesting analysis is how often explicit targets interact with each other. We observe that they are rarely combined, as only *eye for an eye* co-occurs in over 50% of its appearances with others. The two most frequent targets relatively co-occur most with other targets (see column 5 and 15 in Table 1), which underlines their role as proxies for a general position. As both sub-debates contain the demand of the death penalty for certain crimes, unsurprisingly, they also co-occur quite frequent.

While most sub-debates are rarely backed up with others, e.g. *Gunshot*, *Strangulation* or *Electric Chair* are combined frequently (rows 13,14 and 19). We argue that these sub-debates tend to be rather complementary, as they are used to support other sub-debates or form logical units with them. For instance, *Gunshot* frequently co-occurs with *More Quickly*, which is reasonable if one considers that an execution could be performed very quickly by shooting.

Machine Learning In order to inspect the influence of sub-debates on stance detection, we examine how well classifiers perform on comments that contain a certain explicit target. Therefore, we calculate the performance on subsets of the data, which correspond to the instances annotated with the respective explicit target.

To reflect the state-of-the-art, we implement prototypical representatives of the two main strands of approaches – an SVM and a neural architecture with a (Bi)LSTM layer in its core. We calculate the classification performance using leave-one-out cross-validation on the video level. Before executing the classification, we tokenize the data using the ArkTokenizer (Gimpel et al., 2011) from the DKPro Core framework (v1.9.0) (Eckart de Castilho and Gurevych, 2014).

Figure 6 visualizes our results for stance classification. Overall, we achieve an F_1-score of .45 for both classifiers. If we consider only the comments that contain a sub-debate, classification works considerably better (.53 for LSTM and SVM). However, if we exclude comments with sub-debates, we observe a large drop in classification performance (LSTM: .24, SVM: .23). From that, we conclude that classifiers are mainly learning to classify sub-debates.

An interesting special case is the performance on comments that express an *Explicit Debate Stance* towards death penalty. We find that the F_1-score is in the same range (LSTM: .52, SVM: .55) as for the classification of sub-debates. This further supports our decision to treat the explicit debate stance as a special case of a sub-debate.

Overall, we do not observe major differences between the two sub-debate inventories regarding classification performance.

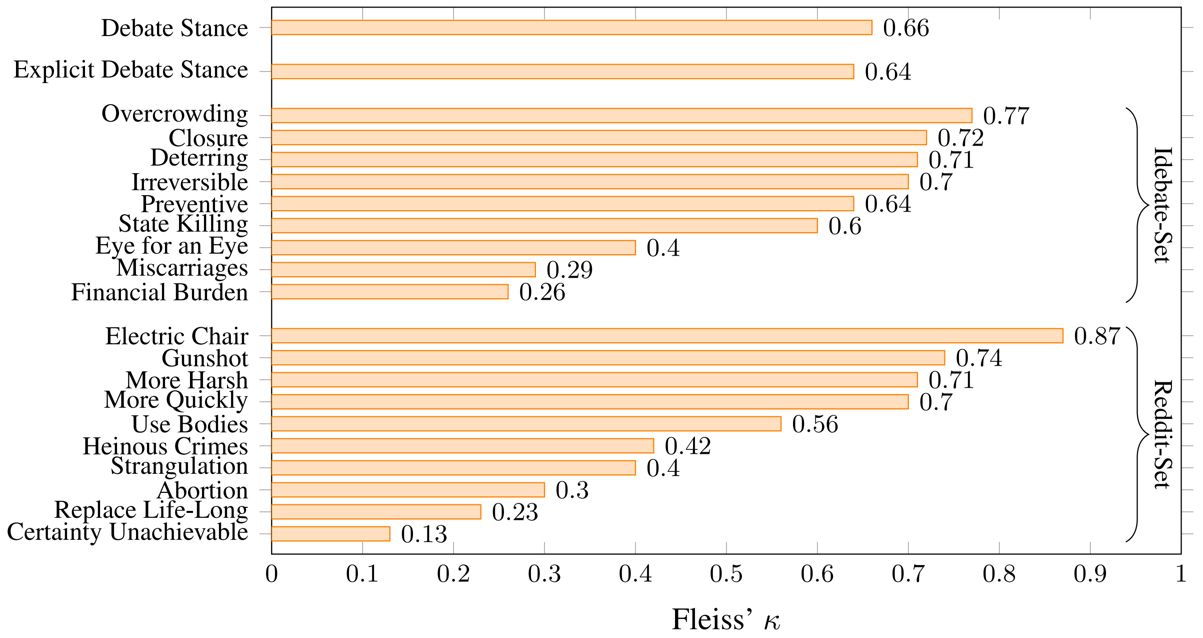

Figure 5: Inter-annotator agreement for the debate stance, explicitly expressed debate stance, the Idebate Set and the Reddit Set.

Sub-Debate		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Explicit Debate Stance	1	**87**	1	8	13	3	0	4	0	5	9	0	9	1	1	15	1	1	2	1	0
Closure	2	1	**19**	3	7	2	1	1	0	3	1	0	1	0	1	6	2	0	1	1	0
Deterring	3	8	3	**55**	5	6	1	2	1	6	3	0	4	0	1	6	2	3	3	0	0
Eye for an Eye	4	13	7	5	**87**	2	1	0	1	6	9	2	5	2	3	45	2	5	6	3	0
Financial Burden	5	3	2	6	2	**46**	2	2	2	2	0	0	7	0	5	6	1	11	4	1	0
Irreversible	6	0	1	1	1	2	**11**	1	0	0	2	0	1	0	0	0	0	0	2	0	0
Miscarriages	7	4	1	2	0	2	1	**19**	0	0	1	0	6	0	0	1	0	0	0	0	0
Overcrowding	8	0	0	1	1	2	0	0	**6**	0	0	0	1	0	1	3	0	0	0	0	0
Preventive	9	5	3	6	6	2	0	0	0	**27**	2	1	2	0	2	8	2	2	1	2	0
State Killing	10	9	1	3	9	0	2	1	0	2	**38**	0	5	0	1	8	0	0	3	1	0
Abortion	11	0	0	0	2	0	0	0	0	1	0	**7**	0	0	0	3	0	1	1	1	0
Certainty Unachievable	12	9	1	4	5	7	1	6	1	2	5	0	**57**	0	2	8	0	7	4	1	0
Electric Chair	13	1	0	0	2	0	0	0	0	0	0	0	0	**6**	1	3	1	0	2	0	0
Gunshot	14	1	1	1	3	5	0	0	1	2	1	0	2	1	**22**	4	1	8	0	4	1
Heinous Crimes	15	15	6	6	45	6	0	1	3	8	8	3	8	3	4	**96**	7	5	7	5	0
More Harsh	16	1	2	2	2	1	0	0	0	2	0	0	0	1	1	7	**20**	0	1	1	1
More Quickly	17	1	0	3	5	11	0	0	0	2	0	1	7	0	8	5	0	**30**	0	2	0
Replace Life-Long	18	2	1	3	6	4	2	0	0	1	3	1	4	2	0	7	1	0	**35**	0	0
Strangulation	19	1	1	0	3	1	0	0	0	2	1	1	1	0	4	5	1	2	0	**12**	1
Use Bodies	20	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	**8**

Table 1: Frequency of sub-debates (bold-faced, underlined cells on main diagonal) and their co-occurrences with other sub-debates. We highlight cases (row-wise) in which a sub-debate frequently co-occurs with another. Light orange for more than 15% overlap, and dark orange for more than 30%.

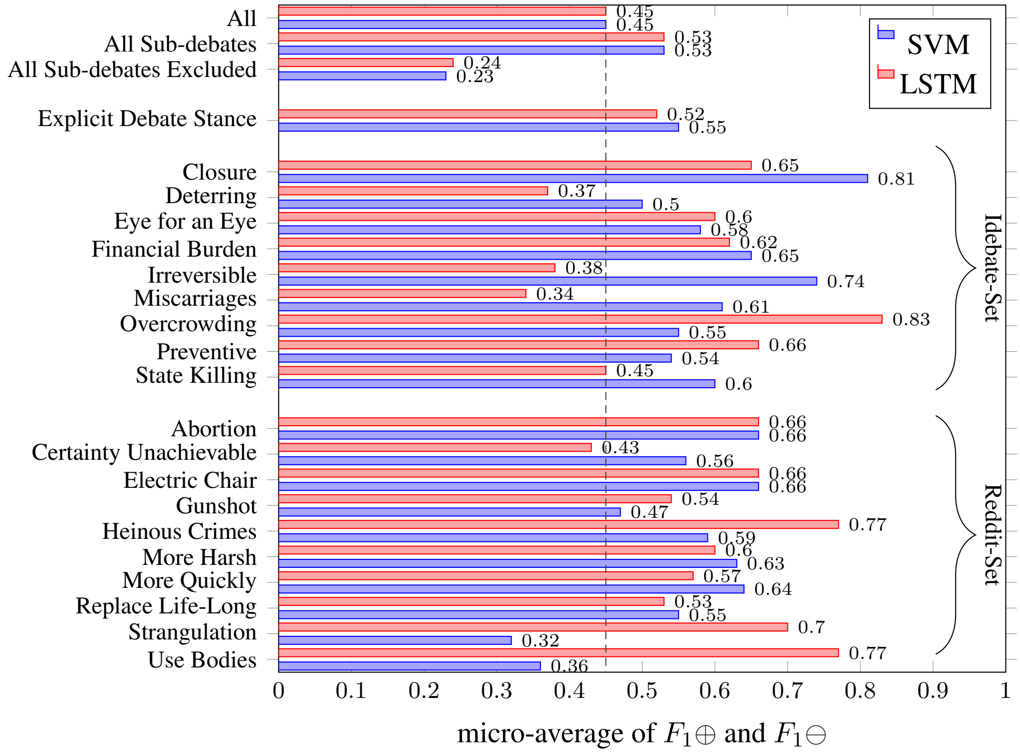

Figure 6: Stance classification performance by sub-debate

5.2 IberEval

For the participating system, we strive to combine the strengths of both strands. Consequently, we first implement prototypical representatives of the strands namely and a neural architecture with a (Bi)LSTM layer in its core and an SVM. Subsequently, we built a third system that automatically decides whether a tweet should be classfied with the neural or the non-neural system. Therefore, we first labeled every tweet with whether the SVM's respectively the LSTM's prediction was wrong or false. We than train an decision tree (weka's J48) for each approach to automate this decision. The trees are equipped with word n-gram coverage, embedding coverage, type token ratio and length features which are suspected of having an influence on the classifiability through the systems. T Unfortunately, the performance of this classification still needs heavy improvement in future work. Based on these classifications we conduct a final decision. In case the system could not derive preference towards one system as both systems are recommend or none, we rely on the SVM as its performance is overall better.

6 Ongoing Activities

6.1 Interactive Stance

6.2 Stancembeddings

7 Future Work

8 Professional Activities

NLP4CMC GermEval2017 IGGSA Reviewer/PC-member for EMNLP, CoNLL, ALW1, BEA

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, Stroudsburg, USA.

Darina Benikova, Michael Wojatzki, and Torsten Zesch. 2017. What does this imply? examining the impact of implicitness on the perception of hate speech. In *to appear*.

Filip Boltužić 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, Baltimore, USA.

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

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.

Adam Faulkner. 2014. Automated classification of stance in student essays: An approach using stance target information and the Wikipedia link-based measure. *Science*, 376(12):86.

James B. Freeman. 1991. *Dialectics and the macrostructure of arguments: A theory of argument structure*, volume 10. Walter de Gruyter.

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

Nancy Green, Kevin Ashley, Diane Litman, Chris Reed, and Vern Walker, editors. 2014. *Proceedings of the First Workshop on Argumentation Mining – Front Matter*. Association for Computational Linguistics, Baltimore, Maryland, June.

Herbert P. Grice. 1970. *Logic and conversation*, volume 3. Academic Press.

Ivan Habernal, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Argumentation mining on the web from information seeking perspective. In *ArgNLP*.

Kazi Saidul Hasan and Vincent Ng. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the IJCNLP*, pages 1348–1356, Nagoya, Japan.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *Proceedings of the EMNLP*, pages 751–762, Doha, Qatar.

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

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Amita Misra, Pranav Anand, JEF Tree, and MA Walker. 2015. Using summarization to discover argument facets in online idealogical dialog. In *Proceedings of the NAACL HLT*, pages 430–440, Denver, USA.

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016a. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the International Workshop on Semantic Evaluation (to appear)*, San Diego, USA.

Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2016b. Stance and sentiment in tweets. *arXiv preprint arXiv:1605.01655*.

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

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. *Proceedings of SemEval 2014*, pages 27–35.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th SemEval*, pages 486–495, Denver, Colorado. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphee De Clercq, Veronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Núria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th SemEval*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Bjrn Ross, Michael Rist, Guillermo Carbonell, Ben Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. In Michael Beiwenger, Michael Wojatzki, and Torsten Zesch, editors, *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, pages 6–9.

Frank Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational linguistics*, 22(1):1–38.

Parinaz Sobhani, Diana Inkpen, and Stan Matwin. 2015. From argumentation mining to stance classification. In *Proceedings of the NAACL HLT 2015*, pages 67–77, Denver, USA.

Dan Sperber and Deirdre Wilson. 1986. Relevance: communication and cognition. *Language in Society*, 17(04):604–609.

Dhanya Sridhar, Lise Getoor, and Marilyn Walker. 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, Baltimore, USA.

Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 217–226, Prague, Czech Republic.

Stephen E. Toulmin. 1958. *The uses of argument*. Cambridge University Press.

Peter D. Turney. 2002. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 417–424. Association for Computational Linguistics.

Marilyn A Walker, Pranav Anand, Robert Abbott, and Ricky Grant. 2012. Stance classification using dialogic properties of persuasion. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 592–596. Association for Computational Linguistics.

Wan Wei, Xiao Zhang, Xuqin Liu, Wei Chen, and Tengjiao Wang. 2016. pkudblab at semeval-2016 task 6: A specific convolutional neural network system for effective stance detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation*, pages 384–388, San Diego, USA.

Michael Wojatzki and Torsten Zesch. 2016a. ltl.uni-due at semeval-2016 task 6: Stance detection in social media using stacked classifiers. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, volume 10, San Diego, USA. ACL.

Michael Wojatzki and Torsten Zesch. 2016b. Stance-based Argument Mining - Modeling Implicit Argumentation Using Stance. In *Proceedings of the KONVENS*, pages 313–322, Bochum, Germany.

Michael Wojatzki and Torsten Zesch. 2017. Neural and non-neural stance classifiers learn sub-debates a study on death penalty debates on youtube. In *to appear*.

Ruifeng Xu, Yu Zhou, Dongyin Wu, Lin Gui, Jiachen Du, and Yun Xue. 2016. Overview of NLPCC Shared Task 4: Stance Detection in Chinese Microblogs. In *International Conference on Computer Processing of Oriental Languages*, pages 907–916. Springer.

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 129–136. Association for Computational Linguistics.

Guido Zarrella and Amy Marsh. 2016. MITRE at semeval-2016 task 6: Transfer learning for stance detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation*, pages 458–463, San Diego, USA.