Mining Implicit Argumentation in Social Media – Status Quo

Michael Wojatzki

Abstract

This status quo document has the following structure: First, we will give a brief overview on the field of Argument Mining, outline problems we identified and state our research questions. Second, we explain the central idea **MW:** wirklich? Afterwards, we will show what we have done so far to examine the quality and usefulness of our model. Finally, we will address remaining questions and problems we see in our approach and the so far conducted experiments.

1 Overview on State-of-the-Art and RQs

MW: why to deal with argmining? **MW:** what is the sota? **MW:** what is the performance of the sota? **MW:** what is the problem? **MW:** How we want to solve them? overal approach: shifting to argument perception!

Argumentation is a constellation of propositions that is used to convince someone of a standpoint. Especially in social media, argumentation is frequently observable and can be considered as an essential element of social media interaction such as online debates. Since this phenomenon occurs at a massive scale, many groups of information seekers (e.g. researchers, journalists, companies, etc.) could benefit from an automated analysis of social media argumentation. This automated identification of argumentative structures within written communication is called argument mining (Green et al., 2014). One yet unsolved problem is that –especially in informal settings – argumentation is often done implicitly. For instance, in a debate on atheism, one may observe an utterance such as *Bible: infidels are going to hell* or even shorter *#JesusOrHell*. In the context of a debate about atheism, both utterances implicitly express the argument that the author is against atheism, because the bible says that this will result in a stay in hell after death. However, both claims are never explicitly mentioned.

Typically, models of argument mining assume that an argument consists of at least an explicit *claim* and a number of optional supporting structures such as *premises* (Palau and Moens, 2009; Peldszus and Stede, 2013).

MW: hier toulmin etc aufdroeseln **MW:** performance overall, paint devastating picture here paper: "AM from info seeker perspective"

After analysing the state-of-the-art, we identified two fundamental problems of current approaches. First, most approaches rely on strict formalisms such as the Claim-Premise-Scheme which requires that an argument is composed of exactly one claim and an arbitrarily large number of premises, justifications and other forms of support (Habernal et al., 2014) . These schemes are developed to model argumentation highly elaborated, well-formed text (e.g. scientific writing or legal text) and less suited to deal with the noise and lower argument density of social media. Second, in contrast to elaborated text, social media contains a high proportion of implicit argumentation (e.g. up to 50% of the claims are implicit in an online debate).

In order to tackle the described problems, we propose stance-based argument mining as the topic for the dissertation. A stance is the attitude (being in favor or against) of an author towards a given target like a politician or a controversial topic (Mohammad et al., 2016a). By transforming propositions into a constellation of stances towards targets, one should obtain a more abstract representation of arguments that should be more robust against implicitness and noise. Therefore, the first milestone for the dissertation is to develop the ability detect the stance towards a defined target. Hence, we have already participated in the *SemEval 2016 Task 6: Detecting Stance in Tweets* with considerable success. As a next step, this ability should be applied to targets which are determined at runtime. Further, experiments will be conducted which will help to model the combination of several targets into a comprehensive

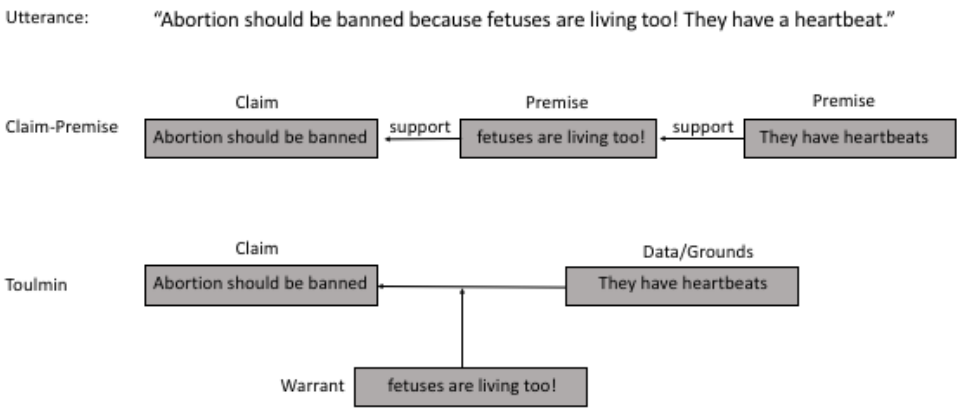

Figure 1: Toulmin and Freeman

schema.

2 Stance-Based Argument Mining

MW: also related work **MW:** overall idea of our approach, details of first scheme in section implicit and explicit our approach

In order to solve the major challenge of implicit arguments that cannot be modelled well with existing approaches, 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 **??** gives on overview of the model which we metaphorically describe as an iceberg. In the context of a debate about atheism, an utterance like *God will judge those infidels!* is like the visible (explicit) part of the iceberg. It expresses a stance in favor of a supernatural power (*Supernatural Power* $\oplus$), while the actual stance on the debate target of atheism (*Atheism* $\ominus$) is not visible but must be inferred. Note that the debate stance might also be explicitly expressed (see figure **??**a), but in implicit argumentation it has to be derived from the explicit stances.

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.

An explicit stance always implicitly covers a large variety of associated stances, we propose the metaphor of an iceberg, whose actual size is also not observable but present and significant under the water surface.

A stance which may be expressed only implicitly and may be inferred from the explicitly made stances is the second component of our model – namely the debate stance. This stance is important for the expressiveness of of our model, since it corresponds to the claim of other models in argument mining.

Note, that the debate stance and other stances that are implicitly covered by the explicit stances are derived as a function of the context of the present debate. They may be a completely different if one states the same utterance in a different debate. For instance, in a debate on whether the Bible is judgmental book the examples in figure**??** would be affirmative to this target. The two layers of the model and the iceberg metaphor are exemplified in figure **??** for an utterance with one explicit stance and one utterance with two explicit stances.

For modeling stance, we can build on plenty of research (Anand et al., 2011; Somasundaran and Wiebe, 2009; Sridhar et al., 2014; Hasan and Ng, 2013) and even a shared task on automatic stance detection (Mohammad et al., 2016a). These works commonly define stance as being in favor of or against a given target. Consequently, stance is a tuple consisting of a target and a stance expression such as *Atheism* $\oplus$ or *Atheism* $\ominus$.

Debates can be categorized in two sided debates

in which authors can take a pro or contra stance and more open debates which may contain several other targets. However, we argue that each of the targets in an open debate can be considered as a two sided debate. Thus, if one acknowledges that the participants in a two sided debate also discuss certain sub-topics, the separation between two sided debates and open debates vanishes.

3 Conducted Experiments and Results

3.1 Automated Stance Detection

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., 2016a), 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; Boltužić and Šnajder, 2014).

Automated stance analysis is closely related to the task of mining arguments or sentiments (Boltužić and Šnajder, 2014). However, in contrast to sentiment analysis, stance is necessarily aimed at a defined target. **TZ:** Das ist zumindest bei aspect-oriented sentiment analysis auch so. Gibt es auch da Unterschiede? It is more general than argument mining, since argument mining deals with finding the reasons why someone has a certain stance or opinion (Boltužić and Šnajder, 2014). The aforementioned SemEval task represents the first shared task on stance detection that tries to explore the current state-of-the-art. **MW:** hier nochmal ins SemEval paper schauen un da die Abgrenzung herauskramen 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.

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 2.

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

For a detailed overview on our features and the implementation see Wojatzki and Zesch (2016a). **MW:** results **MW:** lessons learnt

3.2 Explicit and Implicit Stances

MW: as a first step to utilize stances for a robust argumentation scheme we...

we utilize a semi-automated, bottom-up approach that focusses on concepts that are mostly explicitly expressed by named entities and nouns. Consequently, we examine the frequency distributions of nouns and named entities. Since we observe that the distribution follows Zipf's law, we expect that words with a frequency above the long tail, can serve as candidates as they occur frequently enough to avoid the sparsity problem. It should be noted that in this corpus of Twitter messages on Atheism, the *atheism* appears exactly once and the *atheist* only 6 times. This indicates that implicit argumentation is prevalent in social media.

As we want to ensure that the targets used enable us to differentiate the authors' positions sufficiently, we also consider the degree of association between nouns and named entities to the

[1] version 0.8.0-SNAPSHOT

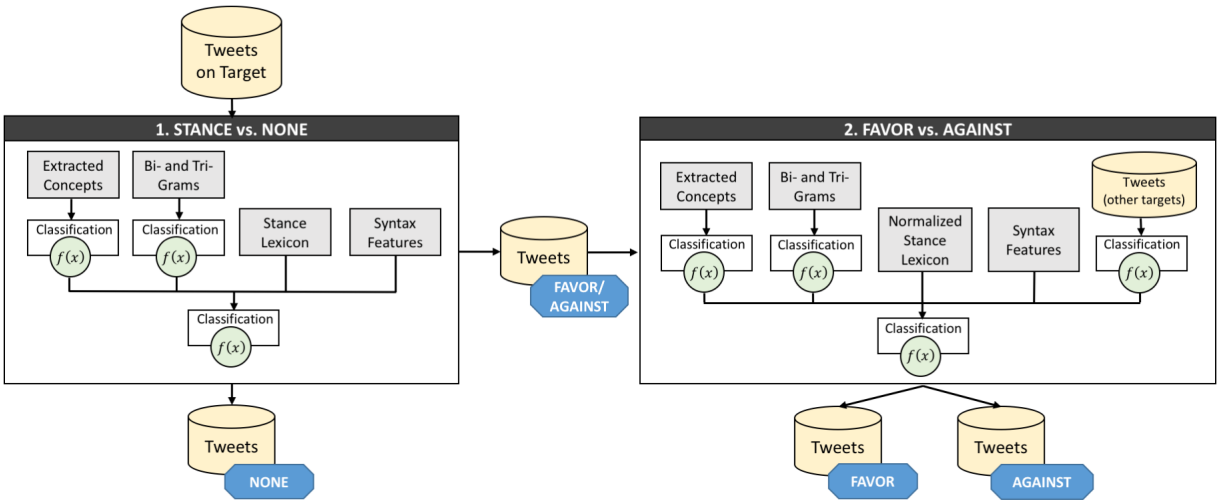

Figure 2: Overview on the sequence of stacked classifications that is used for the supervised setting (subtask A)

stances *Atheism* ⊕ and *Atheism* ⊖. 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* ⊖ and *Atheism* ⊕.

We found the resulting concepts to be too numerous and too fine-grained to be used in our model. We thus, manually group concepts into more coarse-grained targets. For instance, concepts such as *Bible* and *Jesus* are grouped into the target *Christianity*. A potential criticism of our approach is that at this stage of our work, we can not evaluate whether the set is best possible choice. We plan to shed light on this aspect in future research. The final set of derived, explicit targets is shown in table **??**.

3.2.1 Annotation Process

Using the selected data, we let three annotators identify stances towards the derived targets and the debate target. In order to familiarize the annotators with our model, we previously trained them on a small data set that is comparable in its social media character but concerns a different target.

Since the data partly contains utterances which cannot be understood without further context, we give annotators the option to mark them accordingly. Irony is another phenomenon, which influences the interpretability. Therefore, we asked the annotators to annotate the tweets for irony as well.

Since it is still possible that our annotators interpret the tweets differently than in the original annotation, we re-annotated the debate stance using the original questionnaire described in Mohammad et al. (2016a). While annotating explicit stances, the annotators had the instruction to only annotate stances towards targets if they have textual evidence for it.

3.2.2 Evaluation

In this section, we evaluate the annotated data. For this purpose, we first analyze the reliability of the annotation on different levels of granularity using Fleiss' Kappa (κ). For the analysis, we exclude tweets that are annotated for irony and understandability issues. However, we found that the annotators rarely agree on these phenomena as we get a κ of only 0.06 for understandability and a κ of 0.23 for irony. Therefore, we only exclude 18 tweets in which at least two annotators share the same judgment, which results in 715 tweets for the final corpus.

3.3 Inter Annotator Agreement

Since the explicit targets are annotated on the basis of textual evidence, we expect a high level of agreement. The notation of explicit targets should also result in a strong agreement of the annotation of the debate stance because it enforces a deep analysis of the communicative goal of an utterance. As shown in figure 3, we obtain a Fleiss' κ of 0.72 for the annotation of the debate stance. Unfortunately, we cannot compare our agreement to the originally SemEval data, as the organizers do not report a chance corrected agreement measure for their final decision. Also not directly com-

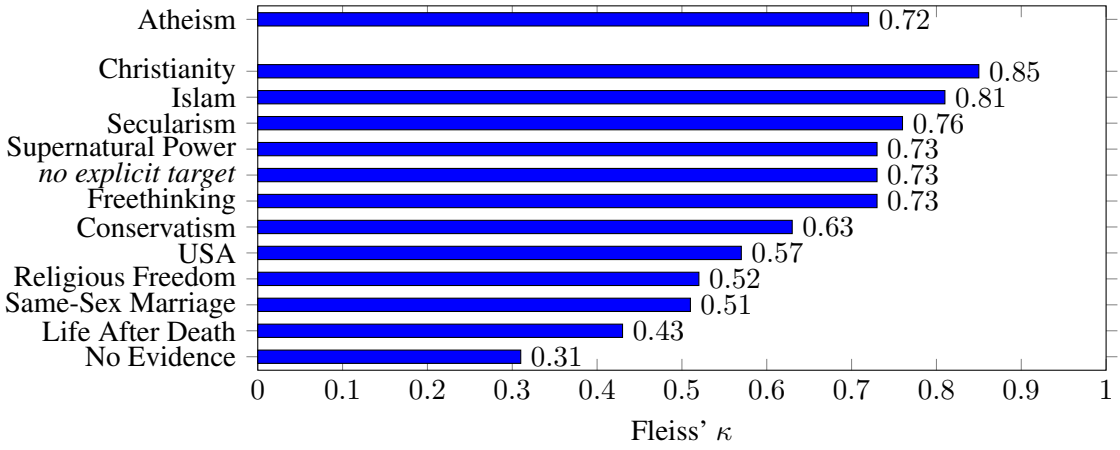

Figure 3: Inter-annotator agreement of the debate stance *Atheism* and explicit stances

parable is the agreement of Sobhani et al. (2015) as they report weighted κ. We argue that their weighted κ of 0.62 is in a range similar to ours.

In figure 3, we also show the agreement for the explicit targets. Since explicit stances have a similar, deriving function like the argumentative phrases proposed by Conrad et al. (2012) and Hasan and Ng (2014), we compare our agreements to theirs which does not exceed a Cohen's κ of 0.68. Two targets (*Christianity* and *Islam*) yield especially high agreement above 0.8, because they are associated with clear signal words such as *Jesus* and *Quran* and other markers such as the numerical reference to biblical passages. Other targets such as *Secularism* and *Freethinking* are rather abstract. They hardly involve special signal words but still gain high agreements of a κ above 0.7, which shows that our annotators did not just learn to recognize certain keywords, but can also reliably annotate more abstract targets. This is further supported by the fact that the agreement for the annotation of *no explicit target* is also in this range. The targets *USA*, *Religious Freedom*, *Same-Sex Marriage*, and *Life After Death* yield only a moderate agreement between 0.4 and 0.6. 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. We therefrom see two possibilities here to improve the agreement: On the one hand, we could choose more comprehensive targets such as *gay rights* to cover the combined positions. On the other hand, we could train the annotators to more consistently account for such differences. A rather low κ of 0.31 is obtained for the target *No Evidence*. Regarding this target, we observe that annotators sometimes deviated from our guidelines and incorporated different degrees of inferred knowledge as they used *Bill Nye* or *Richard Dawkins*[2] as anchors for their decisions, although the utterance contains no explicit stance in favor of *No Evidence*.

Finally, we obtain a κ of 0.63 for the joint decision on both the debate and the explicit targets. Note that this agreement is not directly comparable with the approaches from related work, as they only implicitly model the debate stance, do not report agreements of a joint decision or rely on stances that are determined by the structure of the data. The obtained inter-annotator agreement shows that our model can be annotated reliably and that the recognized difficulties may be compensated by a better training of the annotators and a better selection of targets.

3.3.1 Stance Pattern Analysis

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. This pattern is also observable for the reverse case such

[2] famous supporters of the position that there is no evidence for religion

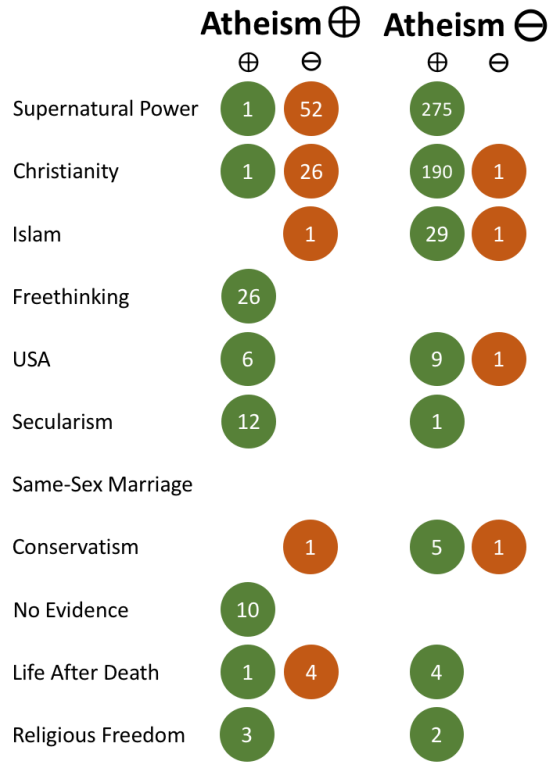

Figure 4: Frequency of explicit stances grouped according to debate stance

as for *Freethinking*. Note that utterances addressing the target *Same-Sex Marriage* are exclusively annotated for expressing no stance towards *Atheism*. Further exceptions are the targets *USA* and *Religious Freedom* that are positively mentioned by both camps. However, a deeper analysis shows that these targets always occur together with other targets which seem to be more relevant for the debate stance.

In order to analyze stance patterns in more details, we show which other stances are used together with the target *Supernatural Power* (the most frequent target in both camps) in figure 5. We observe that authors that are in against Atheism use *Christianity* $\oplus$ together with *Supernatural Power* $\oplus$ in 50% of all cases. In contrast, authors that are in favor of Atheism only combine *Supernatural Power* $\ominus$ with *Christianity* $\ominus$ in 13% of all cases. The figure also shows that the other explicit stances only play a subordinate role in the combination with those targets.

From these analyses 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.

Target (# instances)	Baseline	SVM
Supernatural Power (335)	.53	.78
Christianity (223)	.69	.79
Islam (43)	.94	.95

Table 1: Explicit stance classification (only showing targets occurring in at least 5% of all instances)

3.4 Automatically Assigning Stances

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 (Mohammad et al., 2016a) using the DKPro TC framework[3] (Daxenberger et al., 2014b) and leave the development of sophisticated classification models to future research. For preprocessing, we rely on the DKPro Core framework[4] (Eckart de Castilho and Gurevych, 2014) and apply a twitter-specific tokenizer (Gimpel et al., 2011). In all experiments, we use ten-fold cross-validation and report micro averaged F_1.

Explicit Stances As the result from the stance detection task in SemEval-2016 (Mohammad et al., 2016a) indicate, a support vector machine equipped with simple word and character n-gram features is the state of the art in automated stance prediction. Table 1 shows the results of the state-of-the-art classifier and the majority class baseline for comparison. The results indicate that the two most frequent targets can be classified with success, if one relates them to the majority class baseline. We observe that each target has its own linguistic markers such as the use of Arabic terms if one is in favor of Islam. Therefore, we argue that these peculiarities can be targeted even better by specialized features.

The analysis in table 1 excludes targets that have a insufficient coverage (less than 5% of all instances) to train a meaningful model. A possibility to deal with this sparsity may be to incorporate unlabelled data such as demonstrated for traditional models by Habernal and Gurevych (2015).

[3] version 0.8.0
[4] version 1.7.0

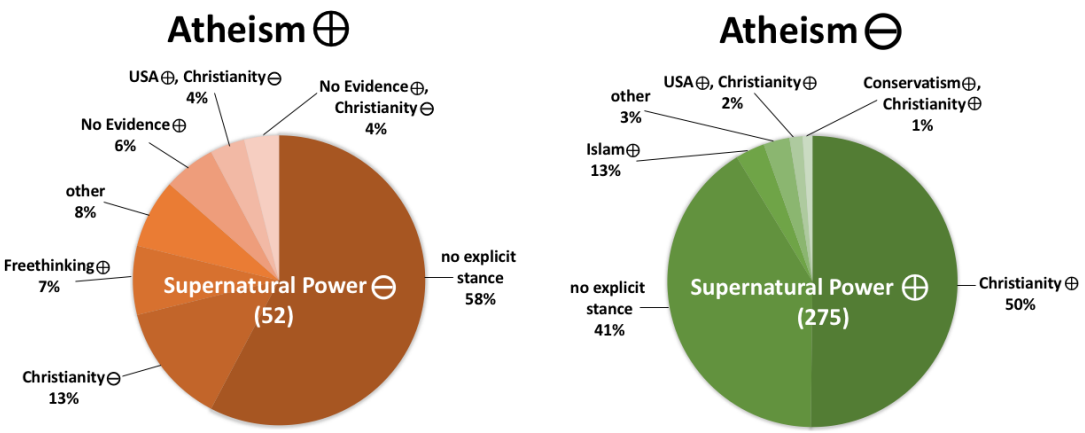

Figure 5: Most frequently used, explicit stances and the percentage shares to which they cooccur with other explicit stances

Feature Set	F_1
baseline	.49
n-gram	.66
ngram + explicit stance$_{predicted}$	.67
ngram + explicit stance$_{oracle}$	.88
explicit stance$_{predicted}$	.65
explicit stance$_{oracle}$	.88

Table 2: Debate stance classification

Debate Stance Table 2 shows the results obtained for automatically assigning the debate stance. 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 for a decision tree classifier (J48). 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 explicit 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.

4 Remaining Problems

The previously conducted experiments are only a first step towards argument mining that is robust against less elaborated and implicit arguments. At the present time we identify four major lines of future work.

4.1 Advanced Machine Learning Techniques and Feature Engineering

So far, we applied standard machine learning techniques (SVM equipped with word n-grams) to the classification tasks. At the moment we modelled the task as a document classification (i.e. the targets are classified independently). Since we assume that there are strong dependencies between the targets, future experiments should implement a sequence classification by using e.g. SVM_{HMM} of conditional random fields.

As indicated by Habernal and Gurevych (2015) and Mohammad et al. (2016b) leveraging huge amounts of unlabelled data is beneficial for stance classification. The idea behind this is to address the data sparsity with approaches such as bootstrapping or distant supervision. In addition, the sparsity problem could be tackled by utilizing lexical semantic resources such as Wikipedia and semantic relatedness of words. Consequently, we want to run experiments that implement these methods.

Moreover our model should profit from features that are tailored to certain explicit targets. For instance, features that capture references pas-

sages by using regular expressions or text similarity measures towards the Bible should be beneficial for classifying the target *Christianity*.

While the debate stance of an utterance is – of course – dependent on the current debate, models for classifying stance towards explicit targets should be domain more domain indepedent. Although this is an assumption that has to be proved, it is hard to imagine a domain in whichI love Jesus the utterance *I love Jesus* does not express a favorization of *Christianity*. Consequently, it should be possible to create a collection of models for classifying explicit stances that could be applied to a new debate in the manner of a building block system.

4.2 Selection of Explicit Targets

In our model, choosing the right number and granularity of targets is crucial. On the one hand, they have to be expressive enough to capture differences in nuanced argumentation. On the other hand, they should not be too fine grained as this would result in very sparse distributions that cannot be handled by automated methods. In our previous work (Wojatzki and Zesch, 2016b), we used a simple frequency approach that focussses on nouns and named entities only. However, at the moment it is unclear how well the resulting set is able to describe the actually used explicit targets. Consequently, we want to examine this problem from three perspectives:

- Gold Standards and Evaluation: Especially critical is to develop appropriate methods that allow to assess the performance of selection approaches. Therefore we need to create gold standard data and choose meaningful measures. We are currently looking for ideas to operationalize such data generation.

- Theoretical Framework: We need a better theoretical understanding of how and what humans perceive as being explicitly expressed. One approach could be to examine the processes which are believed to enable humans to summarize Here we also want to gain a deeper understanding on how this relates to lexical priming effects – the assumed basis of implicit argumentation.

- Modelling and Operationalizability: Ultimately, it comes to find better targets that approximate the created gold data and align theoretical considerations. Therefore we plan to apply more advanced approaches from corpus linguistics such as $tf.idf$ or statistical topic modelling (e.g. LDA). It seems very promising to consider a clustering of synonyms or lexical substitutes of frequent nouns ($tf.idf$ selected nouns, etc.). With this goal in mind we already showed that lexical substitution can model ambiguity well (Wojatzki et al., 2016).

4.3 Overall Model

Insight form the previously conducted experiments suggests that our model can be improved in certain apects. First, our model is adapted to texts that contain a high amount of implicit claims, but cannot express the extent of implicitness. Therefore for future work we want to make sure that the debate target is always contained in the set of explicit targets regardless of how frequent it is explicitly expressed. Second, as the granularity of the explicit targets is still unclear, the best way may be to model copuld be to use multiple layers or explicit targets (increasingly implicit). A question that arises here is whether the model should actually be grounded to the surface form of the utterances or rather to a more abstract representation such as the abstract meaning representation by Banarescu et al. (2012). Of course, such an additional layer may also be a source of error.

Moreover, our current model is adapted to short utterances (i.e. tweets). Although our model allows more than one explicit targets per utterances, it is unclear if it is applicable for long documents. This yields the question on whether our model should be sentence wide or a document-wide one.

4.4 Application

As described above we developed and tested our modle on a narrow domain and a limited amount of data. However, the true value of our model will be seen only if it is applied to different domains and use-cases.

to gain more insight it would also be very interesting what happens if we apply schema to well formed text, formal texts (legal domain) For this purpose, we are always interested in collaborations and interesting application areas for our model :)

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.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2012. Abstract meaning representation (amr) 1.0 specification. In *Parsing on Freebase from Question-Answer Pairs. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle: ACL*, pages 1533–1544.

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.

Johannes Daxenberger, Oliver Ferschke, Iryna Gurevych, and Torsten Zesch. 2014a. 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.

Johannes Daxenberger, Oliver Ferschke, Iryna Gurevych, Torsten Zesch, et al. 2014b. Dkpro tc: A java-based framework for supervised learning experiments on textual data. In *ACL (System Demonstrations)*, pages 61–66, Baltimore, 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.

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.

Ivan Habernal and Iryna Gurevych. 2015. Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse. In *Proceedings of the EMNLP*, pages 2127–2137, Lisbon, Portugal. Association for Computational Linguistics.

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.

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*.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107, New York, USA.

Andreas Peldszus and Manfred Stede. 2013. Ranking the annotators: An agreement study on argumentation structure. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 196–204, Sofia, Bulgaria.

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.

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, Singapore.

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.

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 3rd International Workshop on Argument Mining*, volume 10, Berlin, Germany. ACL.

Michael Wojatzki, Oren Melamoud, and Torsten Zesch. 2016. Bundled gap filling: A new paradigm for unambiguous cloze exercises. In *Proceedings of the Building Educational Applications Workshop at NAACL*, volume 11, San Diego, USA. ACL.