Annotation and Detection of Arguments in Tweets

Robin Schaefer
Applied Computational Linguistics
University of Potsdam
Potsdam, Germany
robin.schaefer@uni-potsdam.de

Manfred Stede
Applied Computational Linguistics
University of Potsdam
Potsdam, Germany
stede@uni-potsdam.de

Abstract

Notwithstanding the increasing role Twitter plays in modern political and social discourse, resources built for conducting argument mining on tweets remain limited. In this paper, we present a new corpus of German tweets annotated for argument components. To the best of our knowledge, this is the first corpus containing not only annotated full tweets but also argumentative spans within tweets. We further report first promising results using supervised classification (F1: 0.82) and sequence labeling (F1: 0.72) approaches.

1 Introduction

In recent years the field of argument mining, which focuses on the automatic identification of argument components and their relations in text, has developed substantially (Stede and Schneider, 2018). However, while the majority of research concentrates on well-structured documents (Moens et al., 2007; Stab and Gurevych, 2014), less work has been done on user-generated web content (Park and Cardie, 2014; Habernal and Gurevych, 2015). This shortcoming poses a problem as systems trained on formal and edited texts tend to be inapt of extracting patterns from the more informal user-generated content (Šnajder, 2016).

In this paper we focus on tweets, which are of great interest for the argument mining community due to the increasing use of the microblogging service Twitter[1] in political online discourse. While some first work on argument mining in tweets exists (Addawood and Bashir, 2016; Dusmanu et al., 2017), only a small number of available annotated corpora have been created that can be utilized for training tweet-specific argument mining systems (Bosc et al., 2016).

To improve on this point, we present a new corpus of German tweets annotated for claim and evidence[2]. To the best of our knowledge, this is the first argument tweet corpus not exclusively annotated with the full tweet as the unit of annotation. Instead, argumentative spans within tweets, henceforth called *argumentative discourse units* (ADU) (Peldszus and Stede, 2013), have been annotated as well. They render the corpus suitable not only for supervised classification but also for sequence labeling approaches. We also present first promising experimental results using this corpus.

This paper is structured as follows: Section 2 gives a short overview of the relevant social media and Twitter-related literature on argument mining. Section 3 describes the corpus, the annotation scheme and the annotation procedure. In Section 4 we present first classification and sequence labeling results using the annotated data. Section 5 discusses our results and gives a brief outlook.

2 Related Work

Related work on tweet-based argument mining has focused on separating argumentative tweets from non-argumentative ones and on defining new Twitter-specific tasks.

[1] https://twitter.com/

[2] We follow Aharoni et al. (2014) and others by using the term *evidence* instead of *premise*.

Proceedings of the 7th Workshop on Argument Mining, pages 53–58
Barcelona, Spain (Online), December 13, 2020.

Addawood and Bashir (2016) present a corpus of English tweets annotated for arguments and evidence types like news media accounts or expert opinions. First, arguments are identified on the full tweet level, followed by the subsequent annotation of evidence types. Annotators achieved Cohen's Kappa scores of 0.67 and 0.79, respectively. An SVM trained on linguistic and Twitter-related features yielded an F1 score of 0.89 on the binary classification task (non-argumentative vs argumentative).

Bosc et al. (2016) describe DART, a Twitter argument corpus annotated for arguments and their relations. In contrast to our work, they do not distinguish claim from evidence but join them in the category *argumentative*. Again, annotations are conducted on the full tweet level and result in a Krippendorff alpha score of 0.81. This corpus is used by Dusmanu et al. (2017) for argument classification. Using a set of lexical, Twitter-specific, semantic and sentiment features, they achieved an F1 score of 0.78 on the binary classification task (non-argumentative vs argumentative). They further investigated approaches to perform fact recognition and source identification.

Wojatzki and Zesch (2016) propose an alternative approach to argument mining in tweets. Specifically, they reconsider the challenging problem of implicit claim detection as a stance classification problem by reformulating implicit claims as implicit stances. This procedure is based on the assumption that an implicit stance can be more easily inferred from the respective tweet. They present the Atheism Stance Corpus, which contains tweets annotated for implicit stances. An SVM trained on token and character n-grams yielded an F1 score of 0.66. Schaefer and Stede (2019) improve on these results using different word and sentence embeddings (F1: 0.78).

Goudas et al. (2014) offer early results for argument mining not specifically on Twitter but on social media. They apply classification to separate non-argumentative from argumentative texts. In a subsequent step, sequence labeling is used to extract ADUs from the latter. This two-step approach makes their work comparable to ours. They report F1 scores of 0.77 and 0.42 for the two tasks, respectively.

3 Corpus Annotation

Our complete initial corpus consists of 77,100 tweets collected in 2019 via the Twitter API using the Python library Tweepy[3]. All tweets contain the keyword *klima* ("climate") and mainly concentrate on the topic of climate change, which was intensely discussed by German media and politics during that time. We conducted the following preprocessing steps.

First, we removed all retweets and excluded non-German tweets using the language identification tool langid (Lui and Baldwin, 2012). These steps led to a subset of 29,525 tweets. In the following, we grouped the tweets into pairs, consisting of a tweet, henceforth called *context tweet*, and the tweet to be annotated, which is a reply to the context tweet and, for this reason, is called *reply tweet*. This approach is motivated by the assumption that tweets in a reply relation are more likely to contain argumentation (Dykes et al., 2020). Moreover, given the short nature of tweets, providing a context is supposed to help interpreting the reply tweet's content. All tweets that were no replies were removed and missing context tweets were collected in an additional step. Finally, we removed all @-mentions at the beginning of a tweet, as these mainly point to the tweet's recipients. The final corpus consists of 12,296 context and reply tweet pairs. For the present study, a subset of 300 tweet pairs was annotated.[4]

3.1 Annotation Scheme

We focus on the two main components of argumentation: claim and evidence. We define a claim as a standpoint towards the topic being discussed (i.e. climate change). In contrast, an evidence unit is a statement used to support or attack such a standpoint. Hence, the crucial difference between claim and evidence is the characteristic of evidence units being always related to another statement while claims can be independent units. We distinguish further between evidence 1) *relating to a claim in the reply tweet*, 2) *relating to a claim in the context tweet* and 3) *relating to claims in both tweets*. Importantly, we do not define an ADU syntactically, e.g. by focusing exclusively on the clause or sentence level.

[3] `https://www.tweepy.org/`

[4] Corpus repository: `https://github.com/RobinSchaefer/climate-tweet-corpus`.

Due to the informal language used in the tweets we consider it appropriate to allow the annotators some flexibility to decide on the actual ADU span.

As distinguishing between claim and evidence can be a quite subjective task, especially on Twitter, annotators were advised to follow our component definitions as close as possible. Statements that function independently of other statements shall be annotated as claims. However, if a statement refers to another proposition either by supporting or attacking it or by giving additional information it shall be annotated as evidence, despite its potential usability as a claim. Therefore, annotators were further instructed to focus on possible causal relationships (in a wide sense) between two statements. If a statement directly follows from another it is likely to be a claim (e.g. [We have to limit CO2 emissions]$_{\text{claim}}$, [as too much CO2 has been shown to increase the greenhouse effect.]$_{\text{evidence}}$). We found that using this strategy to decide on the direction of the argumentation, i.e. which ADU is evidence and which ADU is the claim, facilitated the annotation procedure notably. For our purposes, we do not differentiate between correct and incorrect statements. We also do not explicitly annotate relations between two components.

3.2 Annotation Procedure and Results

Two annotators, one of which is a co-author of this paper, were trained in an iterative two-step procedure. First, both annotators individually labelled a subset of 20 tweet pairs according to the annotation scheme. They compared their results, discussed different interpretations and tried to consolidate them. This procedure was repeated until both annotators felt comfortable in completing the task.

For the actual annotation study we again used a two-step approach. Annotators first had to answer two multiple choice questions asking if a claim or evidence can be identified in the reply tweet. Only if one of the two components was found the annotator would continue to the ADU annotation step. No restrictions on the allowed maximal number of components per tweet were made, as this could potentially have led to differing choices in longer tweets. While annotations themselves only were created for ADU spans, we also derived separate tweet-level annotation sets for claim, evidence and argument (claim or evidence) annotations. Also, we experimented with analysing annotations both on the tweet and the ADU level.

First, we present mean percentages of the ADU annotation frequencies. Of the 300 tweets 14% were annotated as non-argumentative. 27% of the tweets contained exactly one ADU (25%: claim; 2%: evidence). 59% of the tweets were annotated for multiple ADUs (27%: 1 claim & 1 evidence unit; 2% 1 claim & >1 evidence units; 15%: >1 claims & 1 evidence unit) which demonstrates the need for ADU-level annotation even in short texts like tweets.

Metric	Claim	Evidence
Cohen's Kappa	0.55	0.37

Table 1: Inter Annotator Agreement (Questions)

Metric	Multi (s)	Argumentative (s)	Argumentative (t)	Claim (t)	Evidence (t)
Cohen's Kappa	0.38	0.45	0.53	0.55	0.44

Table 2: Inter Annotator Agreement (s = ADU span, t = full tweet)

We calculated Cohen's Kappa scores to measure Inter Annotator Agreement (IAA) (Artstein and Poesio, 2008). As shown in Table 1, results for the claim and evidence questions were 0.55 and 0.37, respectively, which indicates that deciding on the presence of evidence is more subjective. This pattern returns in the scores based on the annotations on the tweet level (Table 2). Whereas results for argument and claim annotations are somewhat similar, the kappa for evidence annotation is reduced. Further, the results show that the multi class annotation (claim vs evidence vs non-argumentative) is particularly difficult. As this task is somewhat subjective in nature, a drop of performance is expected. Although we are aware that the IAA results are relatively low, we consider them acceptable due to the subtlety of the task. This is in line with the interpretation of annotation results by Aharoni et al. (2014), who report 0.39 and 0.4 for claim and evidence annotation tasks, respectively.

4 Experiments and Results

In this section, we present first experimental results based on the annotated corpus. We apply two different approaches: For the tweet-level annotations we trained supervised classification models. This is comparable to the prior studies of Addawood and Bashir (2016) and Dusmanu et al. (2017). In addition, we use the ADU-level annotations for running a sequence labeling approach similar to Goudas et al. (2014). We experimented with different combinations of feature sets, preprocessing steps and models. However, we only present the best results here.

Features	**Target**	**Preproc**	**F1 (w)**	**Precision (w)**	**Recall (w)**
Bigrams	Argument	l,p,s	0.8	0.75	0.86
Pretrained BERT Embeddings	Argument	p	**0.82**	0.8	0.86
Uni- & Bigrams	Claim	l,p	0.79	0.78	0.82
Pretrained BERT Embeddings	Claim	p	**0.82**	0.8	0.85
Uni- & Bigrams	Evidence	l,p	**0.67**	0.68	0.68
Pretrained BERT Embeddings	Evidence	p,s	0.59	0.59	0.62

Table 3: Classification Results (l = lowercase, p = punctuation, s = stopword, w = weighted)

Tweet level. Classification models were trained on different combinations of n-grams and on pretrained BERT-based document embeddings (Devlin et al., 2019). The latter were created using FLAIR, an NLP framework that contains a unified interface for employing different types of text embeddings (Akbik et al., 2019). All shown classification results are yielded using eXtreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016), which is a variant of the Gradient Boosting approach introduced by Friedman (2000). We implemented three different classification tasks based on the respective binary target sets: *argumentative* vs *non-argumentative*, *claim* vs *no claim* or *evidence* vs *no evidence*. All results are 10-fold cross-validated.

Table 3 shows macro F1, precision and recall scores, which are weighted for the unbalanced distribution of classes. Pretrained BERT embeddings yield better F1 scores for argument (0.82 vs 0.8) and claim (0.82 vs 0.79) classifications. Interestingly, a model trained on uni- and bigrams performs better on the evidence task than the BERT-based model (0.67 vs 0.59). Importantly, scores for the argument and claim tasks are substantially higher than for the evidence task.

Features	**Target**	**F1 (w)**	**Precision (w)**	**Recall (w)**
Unigrams	Argument	0.69	0.69	0.72
Linguistic & Twitter Features	Argument	0.7	0.7	0.74
Pretrained BERT Embeddings	Argument	**0.72**	0.73	0.72
Unigrams	Claim	0.53	0.55	0.53
Linguistic & Twitter Features	Claim	0.56	0.58	0.56
Pretrained BERT Embeddings	Claim	**0.59**	0.6	0.59
Unigrams	Evidence	0.73	0.68	0.8
Linguistic & Twitter Features	Evidence	0.73	0.71	0.78
Pretrained BERT Embeddings	Evidence	**0.75**	0.76	0.76

Table 4: Sequence Labeling Results (w = weighted)

ADU level. Sequence labeling models were trained on the following features: 1) unigrams, 2) a combination of linguistic (e.g., n-grams, POS Tags) and Twitter-related (e.g., hashtags, @-mentions) features, 3) pretrained BERT-based word embeddings, which were again created using FLAIR. We chose a Conditional Random Fields approach (Lafferty et al., 2001), using the sklearn-crfsuite[5]. Again, all results

[5] sklearn-crfsuite (`https://sklearn-crfsuite.readthedocs.io`) is a scikit-learn wrapper based on CRFsuite (`http://www.chokkan.org/software/crfsuite/`).

are from 10-fold cross-validation.

In the sequence labeling approach BERT-based models perform best for all three labeling tasks. Using a set of linguistic and Twitter-related features improves the F1 scores compared to the simple unigram models in the argument (0.7 vs 0.69) and claim (0.56 vs 0.53) tasks. However, no improvement is achieved in the evidence task. Interestingly, scores are highest for the evidence task whereas the results for the claim task are considerably lower. This pattern contrasts with the tweet-level classification results.

5 Discussion and Outlook

In this paper we presented a new corpus of German tweets annotated for claim and evidence. While a few previous studies on tweet corpus creation for argument mining exist (Bosc et al., 2016), to the best of our knowledge our corpus is the first tweet dataset with ADU annotations. It is also the first German tweet dataset generally annotated for argumentation.

Although we showed that due to the subtlety of the task relatively low IAA scores were achieved, classification and sequence labeling results based on the dataset are promising. Classifying argument and claim components led to robust F1 scores around 0.8. Solely evidence units posed somewhat of a challenge for the classifier. However, sequence labeling models performed best for evidence units. With both approaches we surpassed the results presented by Goudas et al. (2014).

Given that the IAA scores for evidence annotations were reduced as well, we conclude that evidence units pose an especially hard problem to solve. Recalling our definitions of claim and evidence, this seems intuitive. As evidence units are only defined with respect to claims, a decision has to be made about the exact boundary between both components. Moreover, since tweets tend to contain a high degree of implicitness, it can be demanding to judge if a sequence in fact is relating to a claim. We plan to take this issue into account by refining our annotation scheme further.

Another interesting path of future work will be the continuing development of the argument detector. Following Goudas et al. (2014), one possible way of enhancing results could be building a pipeline based on both classification and sequence labeling approaches. More specifically, a classifier customized for identifying argumentative tweets could function as a filter, thereby allowing to train a sequence labeling model on a purely argumentative tweet set. This could increase the model's precision. To this end, we intend to enlarge the number of annotated data.

Acknowledgements

We would like to thank Polina Krasilnikova for assisting in annotating our data and Crowdee (`https://www.crowdee.com/`) for support with their annotation environment. We further thank the anonymous reviewers for their helpful comments.

References

Aseel Addawood and Masooda Bashir. 2016. "what is your evidence?" a study of controversial topics on social media. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 1–11, Berlin, Germany, August. Association for Computational Linguistics.

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argumentation Mining*, pages 64–68, Baltimore, Maryland, June. Association for Computational Linguistics.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Comput. Linguist.*, 34(4):555–596, December.

Tom Bosc, Elena Cabrio, and Serena Villata. 2016. DART: a dataset of arguments and their relations on twitter. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1258–1263, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Mihai Dusmanu, Elena Cabrio, and Serena Villata. 2017. Argument mining on twitter: Arguments, facts and sources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2317–2322, Copenhagen, Denmark, September. Association for Computational Linguistics.

Natalie Dykes, Stefan Evert, Merlin Göttlinger, Philipp Heinrich, and Lutz Schröder. 2020. Reconstructing arguments from noisy text. *Datenbank-Spektrum*, 20(2):123–129.

Jerome H. Friedman. 2000. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232.

Theodosis Goudas, Christos Louizos, Georgios Petasis, and Vangelis Karkaletsis. 2014. Argument extraction from news, blogs, and social media. In Aristidis Likas, Konstantinos Blekas, and Dimitris Kalles, editors, *Artificial Intelligence: Methods and Applications*, pages 287–299, Cham. Springer International Publishing.

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

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea, July. Association for Computational Linguistics.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, ICAIL '07, page 225–230, New York, NY, USA. Association for Computing Machinery.

Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38, Baltimore, Maryland, June. Association for Computational Linguistics.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *Int. J. Cogn. Inform. Nat. Intell.*, 7(1):1–31, January.

Robin Schaefer and Manfred Stede. 2019. Improving implicit stance classification in tweets using word and sentence embeddings. In Christoph Benzmüller and Heiner Stuckenschmidt, editors, *KI 2019: Advances in Artificial Intelligence*, pages 299–307, Cham. Springer International Publishing.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, Doha, Qatar, October. Association for Computational Linguistics.

Manfred Stede and Jodi Schneider. 2018. *Argumentation Mining*, volume 40 of *Synthesis Lectures in Human Language Technology*. Morgan Claypool.

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

Jan Šnajder. 2016. Social media argumentation mining: The quest for deliberateness in raucousness.