Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6121–6130
Marseille, 20-25 June 2022

GerCCT: An Annotated Corpus for Mining Arguments in German Tweets on Climate Change

Robin Schaefer, Manfred Stede
Applied Computational Linguistics
University of Potsdam
14476 Potsdam, Germany
{robin.schaefer|stede}@uni-potsdam.de

Abstract

While the field of argument mining has grown notably in the last decade, research on the Twitter medium remains relatively understudied. Given the difficulty of mining arguments in tweets, recent work on creating annotated resources mainly utilized simplified annotation schemes that focus on single argument components, i.e., on claim or evidence. In this paper we strive to fill this research gap by presenting GerCCT, a new corpus of German tweets on climate change, which was annotated for a set of different argument components and properties. Additionally, we labelled sarcasm and toxic language to facilitate the development of tools for filtering out non-argumentative content. This, to the best of our knowledge, renders our corpus the first tweet resource annotated for argumentation, sarcasm and toxic language. We show that a comparatively complex annotation scheme can still yield promising inter-annotator agreement. We further present first good supervised classification results yielded by a fine-tuned BERT architecture.

Keywords: Argument Mining, Twitter, Climate Change

1. Introduction

In the last decade the field of argument mining (AM) has developed into a fruitful area of study (Stede and Schneider, 2018; Lawrence and Reed, 2020). AM can be defined as the task of automatically identifying and extracting argumentative structures in natural language. This includes the identification of basic argument components (e.g. claim and evidence) [1] and, optionally, their respective properties (e.g. claim verifiability (Park and Cardie, 2014)). While AM originally had a strong focus on edited texts (Moens et al., 2007; Levy et al., 2014; Stab and Gurevych, 2014), more recently research was extended to the domain of user-generated content, which includes, for instance, debate portals (Al-Khatib et al., 2016a) and social media platforms like Facebook (Bauwelinck and Lefever, 2020) and Twitter (Dusmanu et al., 2017).

With respect to Twitter[2], we argued in Schaefer and Stede (2021) that the platform represents an interesting AM data source for the following reasons. 1) It is frequently used for debating controversial issues, such as climate change (Veltri and Atanasova, 2017); 2) Language on Twitter shows conventions typical for social media, including hashtags and abbreviations, which render the task of AM difficult for models trained on edited text types; 3) Twitter features different conventions of posting tweets, namely single tweets, conversations (a chain of tweets in a reply relation) and threads (a chain of tweets produced by the same Twitter account). This is likely to have an influence on structure and style of the argumentation. In this work, we focus on tweets in a reply relation, as we expect to find interesting argumentation in this interactive scenario.

Since AM on Twitter is generally a demanding task, many approaches so far work with rather simple annotation schemes to model argumentation, often with a special focus on a single argument component, i.e., claim or evidence (Addawood and Bashir, 2016; Bhatti et al., 2021). When moving to annotating different argument components in reply-structure tweets, this increases the difficulty of the task, which may result in comparatively low inter-annotator agreement (IAA) (Schaefer and Stede, 2020).

In this paper we present GerCCT[3], the German Climate Change Tweet Corpus, which consists of 1,200 annotated German tweets collected in 2019 and concentrates on the intensely debated topic of climate change. The tweets are annotated for having different argument properties including, for example, *verifiability* and *reason*. While these rather fine-grained property annotations can already be used for training valuable AM models, we also abstract them into more high-level classes, first to the core components of argumentation (*claim* and *evidence*), and then to the general *argument* (= claim and/or evidence). Thus, the corpus provides three hierarchical layers of argument annotations. Furthermore, we labelled tweets for sarcastic and toxic language in order to enable the development of tools capable of filtering out tweets that fall into these categories. Finally we trained classification models on the annotated corpus and present first promising results.

[1] In line with previous work on AM on Twitter (Addawood and Bashir, 2016; Schaefer and Stede, 2020) we use the terms *claim* and *evidence* instead of *conclusion* and *premise*, respectively.

[2] https://twitter.com/

[3] https://doi.org/10.5281/zenodo.6479492

To summarize, our main contributions to the field of AM on Twitter are as follows.

1. We present a new German tweet corpus, GerCCT, which contains annotations for three layers of argument classes: 1) properties, 2) components, 3) general argument.
2. We additionally annotated the tweets for sarcastic and toxic language. To our knowledge, this is the first AM tweet corpus that is also annotated for these attributes.
3. We trained classification models on the annotated corpus and present first promising results.

The paper is structured as follows. In Section 2, we give an overview of the relevant related work. In Section 3, we describe the annotation scheme, procedure, results and corpus statistics. In Section 4, we present first classification results, before we discuss them in Section 5. We conclude the paper with an outlook in Section 6.

2. Related Work

Related work focuses on AM on Twitter and the detection of different claim and evidence properties. Given the rapidly growing number of papers on AM we will here only discuss the most relevant. For recent comprehensive surveys on AM we refer the reader to Stede and Schneider (2018) and to Lawrence and Reed (2020).

AM on Twitter. The early Twitter-related work of Bosc et al. (2016a) presented DART, an English dataset of about 4,000 tweets annotated for argumentation on the full tweet level. While the authors focused on a single class (+/–)-argumentative (Krippendorff's α: 0.81), without distinguishing claim and evidence, they notably also annotated relations between argumentative tweets (α: 0.67). Bosc et al. (2016b) trained several AM components on this dataset. A Logistic Regression model trained on a set of n-grams and POS tags yielded an F1 score of 0.78 on the argument classification task. Dusmanu et al. (2017) annotated a subset of the DART corpus (#Grexit) and an additional tweet set (#Brexit) for factual vs opinionated content (Cohen's κ: 0.73) and sources (Dice: 0.84; (Dice, 1945)), thereby proposing the two new AM tasks *facts recognition* and *source identification*. A Logistic Regression model yielded a micro F1 score of 0.80. A by-class analysis revealed that the model had particular problems with identifying the factual class. In defining factual content as verifiable information their approach is similar to ours. A difference lies in our more fine-grained approach to evidence categories and our additional focus on the identification of claim and evidence, i.e., both core components of an argument.

Goudas et al. (2014) investigated argument and premise detection using a Greek social media corpus that includes Twitter data. A two-step pipeline was applied. First, argumentative sentences were identified using a classification approach. Second, claims and premise units were detected using sequence labeling. The authors reported F1 scores of 0.77 and 0.42, respectively.

More recently, Bhatti et al. (2021) proposed an interesting approach to AM on Twitter that utilizes hashtags representing a claim (e.g. #StandWithPP ('Planned Parenthood')) and manual premise annotations (Krippendorff's α: 0.79). Best classification results were obtained using a fine-tuned BERT-based approach (Devlin et al., 2019) (F1: 0.71). In comparison to most earlier works on AM on Twitter this approach is characterized by the coverage of full arguments, consisting of claim hashtags and premises. However, one major limitation results from the need of a concrete claim hashtag, which renders the approach only suitable for a subset of discussions on Twitter.

Wührl and Klinger (2021) worked on biomedical Twitter data. The tweets were labelled for containing claims and their being explicit or implicit. In the case of explicit claims, their spans were also annotated. Annotators achieved a Cohen's κ score of 0.56 for binary annotation of claim presence. While challenging, training detection models yielded promising results, especially for explicit claims.

Claim and evidence properties. Important work on claim properties was presented by Park and Cardie (2014), who annotated claims in user comments with respect to their verifiability. Also, verifiable claims were separated into non-experiential and experiential propositions, the latter of which are statements referring to a person's experience. Annotators achieved a Cohen's κ score of 0.73. Furthermore, the authors stated that different claim types require different types of support as well. While a verifiable non-experiential claim needs to be supported via objective evidence, an unverifiable claim can only be supported by providing reasons. While we do not differentiate between experiential and non-experiential claims, we still utilize the concept of verifiability for our work. Also, we distinguish between subjective reasons and more objective evidence.

In addition to Park and Cardie (2014), other work highlighted the relevance of evidence properties, often called evidence types, as well. For instance, Aharoni et al. (2014) proposed an annotation scheme based on a claim and three types of evidence: study, expert and anecdotal. Annotating these classes in Wikipedia articles yielded Cohen's κ scores of 0.39 and 0.40 for claim and evidence annotation, respectively. In a similar vein, Al-Khatib et al. (2016b) suggested the evidence properties statistics, testimony and anecdote. This work differs from Aharoni et al. (2014) by using editorials as a data source and by having a special focus on argument strategies. An overall Fleiss' κ score of 0.56 was achieved. Furthermore, Addawood and Bashir (2016) annotated different evidence types in tweets, including news, expert opinion and blog posts

	Tweet Examples	Annotations
1)	Such random prices render the public transport unappealing and expensive. [...] [link]	UC EE
2)	You cannot negotiate with nature. This is why you cannot prepare a climate protection package like a trade agreement. It's about science and its laws are non-negotiable. [...]	reason UC VC
3)	The biggest issue for the climate are people. There are calculations based on the assumption of 50 tonnes of CO2 per person. The planet is suffering from overpopulation. To not get kids is the best you can do for the environment. [...]	UC VC reason
4)	It already starts with the definitions. [...] What is climate change denial? I personally don't know anyone who doubts the human influence.	UC IE
5)	If the climate was a bank, they would have saved it long ago. [...]	sarcasm
6)	You are a criminal and anti-constitutional organisation and know as much about climate as a pig knows about backstroking!	toxic

Table 1: Tweet examples with annotations. Tweets are cleaned and translated from German (UC=Unverifiable Claim; VC=Verifiable Claim; EE=External Evidence; IE=Internal Evidence).

(Cohen's κ: 0.79). An SVM model trained on a set of n-grams, psychometric, linguistic and Twitter-related features yielded an F1 score of 0.79. While these works are similar to our approach, we separate the proposed evidence properties into the categories of *external* and *internal evidence*.

3. Corpus Annotation

In our work we annotated tweets from a collection of German tweets on climate change that we had initially described in Schaefer and Stede (2020). The tweets are arranged in pairs consisting of a reply tweet and its respective source tweet. The latter is included as additional context to facilitate the interpretation of the reply tweet, but only the reply tweets are being annotated.[4] While the initial experiment had been on 300 tweets only, the corpus we are now releasing is quadrupled in size. Besides, we improve on the previous work by applying a more fine-grained annotation scheme that focuses in particular on argument properties. All our annotations are conducted on the document level, i.e. as attributes of complete tweets.

3.1. Annotation Scheme

In Schaefer and Stede (2020, pg.54), we define a claim as "a standpoint towards the topic being discussed" and an evidence unit as "a statement used to support or attack such a standpoint". This definition places a strong weight on the correct annotation of claims given that the subsequent evidence annotation depends on it. In other words, if annotators identify different segments as claims this likely has consequences for the identification of evidence units as well.

[4]While argumentation can certainly unfold across a chain of more than two tweets, we decided on focusing on pairs to facilitate the annotation task. Investigating more complex Twitter conversations represents an interesting future task.

To mitigate this potential source of disagreement, we now apply an annotation scheme that comprises a more differentiated approach to claim and evidence annotation. In particular, the scheme allows for labeling different evidence properties, some of which do not require the existence of a claim. In addition, we label sarcasm and toxic language independently of the argumentative categories. See Table 1 for example tweets with their respective annotations.

Our new annotation scheme is based on the guidelines Wilms et al. (2021) used for annotating German Facebook comments for the *GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments* (Risch et al., 2021). Though they were not developed specifically for argument annotation, we consider those guidelines as a suitable starting point for annotating argumentation in tweets. Our corpus is annotated with the following categories.

- Argument Component
  - Claim
    - Unverifiable Claim
    - Verifiable Claim
  - Evidence
    - Reason
    - External Evidence
    - Internal Evidence
- Sarcasm
- Toxic Language

The claim and evidence properties (*verifiability*, *reason*, etc.), as well as *sarcasm* and *toxic language* are explicitly annotated, whereas the classes *claim*, *evidence* and *argument* are automatically derived from the annotated properties and added to the annotations. *Argument* merely distinguishes argumentative from non-argumentative tweets. In the following we describe the

annotation categories in detail and point out certain differences to the original GermEval scheme proposed by Wilms et al. (2021).

Claim. Following Park and Cardie (2014) we distinguish unverifiable and verifiable claims. Both subcategories are represented in the annotation scheme of Wilms et al. (2021) and called *opinion* and *fact-claiming statement*, respectively. In line with Wilms et al. (2021) we define an unverifiable claim as a subjective standpoint, positioning, interpretation or prognosis. Although such a statement is unverifiable it can still be sufficiently supported by providing reasons.
A statement is classified as verifiable if it can potentially be verified via an external source. Crucially, presenting a statement as verifiable by using linguistic markers alone is not sufficient. This deviates slightly from the GermEval guidelines which define their *fact-claiming* category less restrictive than our definition of verifiability. Potential sources for verifiable claims include, for example, scientific references, statistics, political manifestos and lexicon entries. Importantly, verifiability does not imply factual correctness.
While unverifiable and verifiable claims are mutually exclusive, a tweet can still contain both claim types. Further, claims do not require the occurrence of an evidence unit to be treated as an argument component. This accounts for the often incomplete argument structure in tweets.

Evidence. Contrasting with Park and Cardie (2014), we do not consider different evidence types for unverifiable and verifiable claims. Instead we annotate three general types of evidence: *reason*, *external evidence*, and *internal evidence*. Only a reason is necessarily related to a certain unverifiable or verifiable claim. It is defined as a proposition justifying a claim. As such it depends on an often causal relation between claim and evidence. This implies that a reason can only occur in tweets that contain a claim. Importantly, annotators were told to prioritize claims over reasons if doubts remain with respect to the latter.
In contrast, the occurrence of a claim is optional for the annotation of *external evidence* and *internal evidence*. We decided to modify the GermEval guidelines in this respect for two reasons: 1) As claims tend to be uttered without evidence, evidence is often given independently of an explicit claim; 2) We expect positive effects on the IAA, as evidence can still be reliably identified in cases where annotators disagree on the occurrence of a claim.
We define *external evidence* as a source of proof for an explicit or implicit claim. As verifiability, *proof* does not imply factual correctness. Instead, the source is merely offered as evidence and may need additional fact-checking. This, however, is beyond the scope of this work. While there is a conceptual overlap with the aforementioned sources for verifiable claims, the notions are not synonymous. External evidence is defined as actually provided evidence, whereas the concept of potential sources for verifiable claims is only utilized to judge their verifiability. External evidence includes News, expert opinions, blog entries, books, petitions, images and quotations. Note that external references are often inserted via links. Hence, we treat them as external evidence.
Finally, *internal evidence* represents the author's personal experiences and insights, which includes experiences of their social environment, such as family members.

Annotation Class	Krippendorff's α
Unverifiable Claim	0.63
Verifiable Claim	0.64
Reason	0.41
External Evidence	0.83
Internal Evidence	0.40
Argument	0.71
Claim	0.69
Evidence	0.64
Sarcasm	0.46
Toxic Language	0.69

Table 2: Inter-annotator agreement.

Sarcasm and Toxic Language. AM on Twitter is often made complicated by a substantial amount of sarcastic and toxic language. Contrasting with the GermEval guidelines, which treat sarcasm as an instance of toxic language, we decided to treat the two concepts as individual classes. We define *sarcasm* as humorous criticism often used in combination with irony, while *toxic language* includes vulgar and insulting words and serves the purpose of degrading and discriminating others. Toxic language and sarcasm are not mutually exclusive, as the latter can also be used to degrade. Crucially, while a sarcastic tweet can also contain non-sarcastic and potentially argumentative segments, a toxic tweet is always treated as non-argumentative.

3.2. Annotation Procedure

Two annotators, one of which is a co-author of this paper, were trained to perform the annotation task, which consists of three subtasks for a tweet: 1) Identify toxic language; 2) identify sarcasm; 3) if the tweet is not labelled as toxic and contains non-sarcastic segments, annotate the argument component attributes.
Annotator training took place as an iterative process. Both annotators labelled a subset of 30 tweets according to the scheme and discussed their decisions in a following mediation session, in order to gain familiarity with the scheme and solve open questions. This procedure was repeated thrice.
Once annotators were able to solve the task, a set of 300 tweets was annotated in batches of 100 tweets in order to evaluate the annotation scheme. After each batch we

Set	UC		VC		Reason		EE		IE		Sarcasm		Toxic	
A1	186	0.62	65	0.22	29	0.10	48	0.16	3	0.01	56	0.19	49	0.16
A2	177	0.59	78	0.26	23	0.08	43	0.14	2	0.01	33	0.11	51	0.17
Full Corpus	703	0.59	244	0.20	132	0.11	165	0.14	11	0.01	204	0.17	173	0.14

Table 3: Absolute occurrences (left column) and proportions (right column) of argument properties, sarcasm and toxic language classes calculated for the IAA sets of both annotators (A1 & A2) and the full corpus (n=1,200). The full corpus includes the A1 set (UC=Unverifiable Claim; VC=Verifiable Claim; EE=External Evidence; IE=Internal Evidence).

Set	Argument		Claim		Evidence	
A1	219	0.73	205	0.68	77	0.26
A2	211	0.70	203	0.68	67	0.22
Full Corpus	844	0.70	784	0.65	295	0.25

Table 4: Absolute occurrences (left column) and proportions (right column) of argument, claim and evidence classes calculated for the IAA sets of both annotators (A1 & A2) and the full corpus (n=1,200). The full corpus includes the A1 set.

monitored the IAA. Due to declining IAA for verifiable and unverifiable claims after 200 tweets, we utilized the last batch to refine our annotation scheme further. Afterwards, both annotators revised their annotations of the 200 already labelled tweets and annotated a last so far unseen batch of 100 tweets according to the refined annotation scheme. All IAA scores presented in this paper are based on these 300 annotations. Once the annotation scheme was validated the annotators continued to individually label additional tweets until the current corpus size of 1,200 tweets was reached.

3.3. Annotation Results

We evaluate the IAA (see Table 2) in terms of Krippendorff's α (Artstein and Poesio, 2008).

To begin with, annotating the claim properties *unverifiable* and *verifiable* yields promising IAA of 0.63 and 0.64, respectively. Importantly, while the granularity of the annotation scheme is designed to facilitate clear decisions, the task of annotating argumentation remains quite subjective, which is reflected in the comparatively low IAA of *reason* (0.41) and *internal evidence* (0.40). However, annotating *external evidence* yields a high α score of 0.83.

In addition to the annotations of argument properties, we also calculated IAA scores for the derived claim and evidence components and the argument class (claim and/or evidence). *Argument* yields an IAA of 0.71, and it is notable that *claim* obtains a higher score (0.69) than its properties *unverifiable* and *verifiable*. *Evidence* shows a satisfactory score of 0.64 despite the notable difference between *external evidence* and *reason/internal evidence*. The IAA scores of argument, claim and evidence are substantially higher than the scores we presented in our earlier study (Schaefer and Stede, 2020), which had obtained Cohen's κ scores of 0.53, 0.55, and 0.44, respectively.

The IAA scores of the non-argumentative classes *sarcasm* and *toxic language* showed mixed results. While annotators were able to reliably identify toxic tweets (0.69), sarcastic tweets seem to have been more demanding (0.46).

3.4. Corpus Statistics

We utilized the SoMaJo tokenizer (Proisl and Uhrig, 2016) to calculate basic corpus size statistics based on the reply tweets. Tweets consist of 1-62 word tokens with a mean word count of 32. Tweets containing only one token (n=11) usually hold links. Further, tweets consist of 1-8 sentences with a mean sentence count of 2. In total the corpus consists of 38,350 tokens and 2,850 sentences.[5]

Calculating proportions of annotations reveals substantial class imbalance. Proportions for both IAA annotation sets and the full annotated corpus (n=1,200; including IAA annotations of annotator I) are given in Tables 3 and 4. We will only describe proportions of the full corpus.

Table 4 shows that 70% of tweets were labelled as argumentative. 65% of tweets contained at least one claim and 25% contained at least one evidence unit. The proportions in Table 3 show that unverifiable claims (59%) were more often identified than verifiable ones (20%). With respect to evidence, tweets were most frequently annotated as containing external evidence (14%). Importantly, internal evidence was rarely found in the dataset (1%), while 11% of tweets contained a reason. Sarcasm and toxic language were found in 17% and 14% of tweets, respectively.

Calculating corpus-wide co-occurrence proportions for argument properties (see middle section of Table 5) reveals interesting findings. 14% of tweets contain both unverifiable and verifiable claims. Moreover, reasons

[5]Note that while SoMaJo performed best in the *EmpiriST 2015 Shared Task on Automatic Linguistic Annotation of Computer-Mediated Communication and Web Corpora* (Beißwenger et al., 2016), our tweets somewhat posed a challenge for the tool's sentence segmentation. Hence, we consider the number of tokens a more accurate measure of corpus size.

	UC	VC	Reason	EE	IE
UC	703	163	123	86	10
VC	163	244	34	52	3
Reason	123	34	132	10	2
EE	86	52	10	165	1
IE	10	3	2	1	11
Single	387	54	0	59	1
UC	0.59	0.14	0.10	0.07	0.01
VC	0.14	0.20	0.03	0.04	0.
Reason	0.10	0.03	0.11	0.01	0.
EE	0.07	0.04	0.01	0.14	0.
IE	0.01	0.	0.	0.	0.01
Single	0.32	0.04	0.	0.05	0.
UC	1.	0.23	0.17	0.12	0.01
VC	0.67	1.	0.14	0.21	0.01
Reason	0.93	0.26	1.	0.08	0.02
EE	0.52	0.32	0.06	1.	0.01
IE	0.91	0.27	0.18	0.09	1.

Table 5: Co-occurrence matrices of argument properties (UC=Unverifiable Claim; VC=Verifiable Claim; EE=External Evidence; IE=Internal Evidence). Top: absolute co-occurrences; Middle: proportions (whole corpus); Bottom: proportions (calculated row-wise with respect to given class). The "single" rows show counts/proportions of tweets where only the respective class was annotated.

co-occur more often with unverifiable (10%) than with verifiable claims (3%). This pattern also holds for external evidence, although by a smaller difference (7% vs 4%). The table further shows that about 40% of tweets only contain one argument property. This is especially the case for unverifiable claims, which were exclusively annotated in 32% of the tweets. This, however, does not imply that these tweets only contain one argument component, because a tweet can contain several components with the same property. Calculating proportions for argument components shows that 20% of tweets contain both claim and evidence units.

The bottom section of Table 5 shows proportions that were calculated row-wise with respect to the respective argument property. While the basic patterns of the corpus-wide proportions are confirmed, some new observations can now be made. For instance, the vast majority of reasons co-occurs with unverifiable claims (93%), compared to a substantially lower proportion that co-occurs with verifiable claims (26%). The same tendency holds for internal evidence (91% vs 27%) and somewhat less prominently also for external evidence (52% vs 32%). However, a larger proportion of verifiable (21%) than unverifiable claims (12%) co-occurs with external evidence.

4. Classification

In this section, we present first results yielded by different classification models that we trained on our annotated data.

4.1. Approaches

In order to solve the different classification tasks at hand, we experimented with variations of feature sets and classification algorithms. Feature sets include n-grams and BERT document embeddings, while classification algorithms include, for example, XGBoost, Logistic Regression, Softmax, Naive Bayes and Support Vector Machines. XGBoost models were trained using the package proposed by Chen and Guestrin (2016), while Softmax classifiers were trained using Flair (Akbik et al., 2019). All other models were trained using scikit-learn (Pedregosa et al., 2011). Importantly, we did not aim at optimizing our models, but instead show first results that can be achieved using our annotated data. Hence, we mainly applied the packages' default settings for hyperparameters. Only the results of the most successful classifiers per feature set are presented in this paper.

Our first approach is based on simple unigrams in combination with an XGBoost classifier. Experiments with bi- and trigrams did not yield better results. For preprocessing, newline characters were removed and all links were normalized by replacing them with a placeholder ([link]). The latter is assumed to be beneficial especially for the task of (external) evidence detection, which crucially depends on the correct identification of links. In addition, all tweets were lowercased and punctuation[6] was removed. Removing stop words did not improve results, so we eventually did not apply this step.

We also experimented with two Transformer-based approaches that rely on pretrained BERT document embeddings. These approaches were implemented using Flair, an NLP framework which provides simple interfaces to utilize and train Transformer architectures. We tried different German BERT models and decided on using the *bert-base-german-cased* model by deepset[7] due to its better results. In the first approach we embed our tweets and use the embedding features as input for the classification algorithms, which we also trained on unigrams. Thus, the pretrained embeddings are *frozen*, i.e. not fine-tuned during training, which renders this BERT approach comparable to the more simplistic unigram approach. However, here we utilize a Logistic Regression algorithm. The second BERT approach, in contrast, relies on fine-tuning the BERT architecture. Improved results can be expected from this procedure, because the original BERT model was not trained on social media data, but on Wikipedia, Open Legal Data, and news articles. For classification this pipeline utilizes a Softmax classifier. Fine-tuning of the BERT architecture and training of the Softmax classifier took

[6]The exception are #, and square brackets. Hashtags are frequently used in Twitter data, while square brackets were kept because of the used link placeholder.

[7]`https://www.deepset.ai/german-bert`

place in joint fashion. In this Flair-based approach we trained models for 10 epochs using a learning rate of 5e-05 and a batch size of 4.

Although the classes are imbalanced in our corpus, we refrained from applying under- or oversampling techniques, in order to present classification results that are in general expectable from a Twitter dataset on a controversial topic. For training and testing the data was split in a stratified manner. All presented results stem from 10-fold cross-validation, and they are macro F1 scores. All approaches are cross-validated using the same train-test splits. However, we additionally utilized a 10% subset of the training data for validating the training process in our second BERT approach which relies on fine-tuning. This differs from our other two approaches, which do not require validation during training.

4.2. Results

We compare all classification approaches to a majority baseline, i.e., to a model that naively outputs the majority class for all data instances. Importantly, all approaches substantially outperform this baseline. Note that we do not present results obtained by models trained on *internal evidence*, because this class is extremely rare in the corpus. The few cases of *internal evidence* were also excluded from the dataset used for evidence detection.

Table 6 shows results from models that were trained on the annotations of the fine-grained argument properties, sarcasm and toxic language. It turns out that the unigram approach yields decent results, though there are notable differences between the classes. Unverifiable claims are more successfully detected than verifiable claims (0.63 vs 0.56). Also, F1 scores of the evidence types reason and external evidence differ substantially (0.54 vs 0.81), while in comparison sarcasm and toxic classification show similar results (0.58 vs 0.54).

Applying pretrained BERT embeddings without fine-tuning improves on the unigram results. Most of the previously observed patterns are maintained. Unverifiable claims are still more successfully detected than verifiable claims, though by a smaller difference (0.66 vs 0.62). External evidence units still show the highest F1 score (0.84) and clearly outperform the reason class (0.55). However, contrasting with the unigram approach the BERT approach identifies toxic language more robustly than sarcasm (0.68 vs 0.62).

Fine-tuning BERT embeddings during training has a positive effect on the results of some classes. For instance, the claim properties *unverifiable* and *verifiable* benefit from additional fine-tuning (unverifiable: 0.70 vs 0.66; verifiable: 0.69 vs 0.62), as does the reason class (0.60 vs 0.55). However, external evidence and toxic language only show small variation compared to the scores obtained from the *frozen* BERT architecture (external evidence: 0.86 vs 0.84; toxic: 0.66 vs 0.68). Finally, fine-tuning appears not to enable the BERT model to successfully capture sarcasm (0.48 vs 0.62).

Classification results obtained from models trained on the coarse-grained argument component annotations can be found in Table 7. While all classes yield promising results, evidence detection shows higher scores than claim detection for both the unigram approach (0.72 vs 0.63) and BERT approaches (no fine-tuning: 0.74 vs 0.68; fine-tuning: 0.77 vs 0.73). Fine-tuning the BERT architecture improves the results for both detection tasks. The models trained on the general *argument* class also show good results. Here, however, the BERT architecture benefits from fine-tuning only slightly (0.70 vs 0.69).

5. Discussion

Annotating argumentation especially in user-generated data like tweets is a rather subjective task. While we achieved promising IAA scores for most classes, some of them appear to be more difficult to annotate, which can indicate a higher degree of subjectivity. We will discuss them in turn.

With a Krippendorff's α of 0.41, the reason class proved to be a particular challenge for annotators, especially if compared to the high agreement of 0.83 achieved for external evidence. Recall that reason is defined as a supporting statement which is directly related to a claim. As we suggested in Schaefer and Stede (2020), the close connection between evidence and claim may increase the difficulty of the annotation task. Moreover, while the causal link between a reason and a claim may be explicitly marked by a connective, it tends to be left implicit. As annotators were advised to refrain from annotating reasons if in doubt, the implicitness of a reason-signalling marker may have led to annotations of claims instead. In contrast, external evidence tends to be marked by links, quotations or explicitly named external sources (such as experts), which supposedly contributes to the high IAA for external evidence.

As for the reason class, the Krippendorff's α of internal evidence leaves room for improvement (0.40). One contributing factor is likely the rare occurrence (1%) of internal evidence in the corpus. Also, both internal evidence and unverifiable claims tend to show similar linguistic markers, e.g. 1st person pronouns, as they describe personal positioning and interpretation. Hence, the task of differentiating between internal evidence and unverifiable claims is not trivial.

Finally, the IAA score of sarcasm annotation (0.46) indicates that annotators have difficulty with labelling this particular class. Compared to toxic language which tends to be characterized by the use of explicit degrading vocabulary, sarcasm may be more subtle from a linguistic perspective. While language offers certain indicators, e.g., the winking face emoticon, they are often left implicit, thereby contributing to the challenge of reliably differentiating between literal and sarcastic readings of the same text. Thus, annotating sarcasm

Approach	UC	VC	Reason	EE	Sarcasm	Toxic
Majority	0.37	0.44	0.47	0.46	0.45	0.46
Unigrams + XGBoost	0.63	0.56	0.54	0.81	0.58	0.54
BERT + LR	0.66	0.62	0.55	0.84	**0.62**	**0.68**
BERT (ft) + Softmax	**0.70**	**0.69**	**0.60**	**0.86**	0.48	0.66

Table 6: Classification results for argument properties and sarcasm and toxic language. All scores are macro F1 scores (UC=Unverifiable Claim; VC=Verifiable Claim; EE=External Evidence; ft=fine-tuning; LR=Logistic Regression).

Approach	Argument	Claim	Evidence
Majority	0.41	0.40	0.43
Unigram + XGBoost	0.66	0.63	0.72
BERT + LR	0.69	0.68	0.74
BERT (ft) + Softmax	**0.70**	**0.73**	**0.77**

Table 7: Classification results for argument components. All scores are macro F1 scores (ft=fine-tuning; LR=Logistic Regression).

remains a challenging task.

While IAA scores of these classes indeed leave room for improvement the substantial agreement on the other classes suggests the suitability of our scheme for the annotation of argument properties in tweets. We also present substantially higher IAA compared to the annotation results reported in Schaefer and Stede (2020), which indicates that a more fine-grained annotation approach with respect to claim and evidence components can be beneficial to IAA.

The class distribution shows a high proportion of argumentative content in the corpus: 70% of the annotated tweets contain at least one type of argument component. Given that climate change is a controversial topic, this does not come as a surprise. Also, more tweets contain claims (65%) than evidence units (25%), which shows that claims on Twitter are frequently left unsupported. The majority of claims is unverifiable, which indicates that users contributed to the discourse on climate change more often by stating their opinionated views instead of providing verifiable statements. Moreover, 14% of tweets contain toxic language, which indicates the importance of considering this issue when creating annotated tweet resources.

According to the co-occurence proportions, reasons are predominantly used to support unverifiable claims while external evidence often appears in conjunction with verifiable claims. This is an interesting finding given that our annotation scheme did not require this distinction. It also lends support to the approach adopted by Park and Cardie (2014).

With respect to classification we achieve promising results for most classes. Fine-tuned BERT embeddings in combination with a Softmax classifier yield consistently good results on those annotation classes that show substantial IAA. In contrast, models trained on reason and sarcasm annotations, which show the lowest IAA, also yield the lowest F1 scores. This emphasizes the importance of a well designed annotation scheme that reduces the difficulty of the task.

6. Conclusion & Outlook

In this paper, we presented the German Climate Change Tweet Corpus, which consists of 1,200 tweets annotated for different argument component types and their properties, sarcasm and toxic language. We showed that IAA is promising for the majority of classes, and we presented first good classification results.

This corpus benefits from the annotation of different layers of argument classes. AM tools can be trained that either adopt a more coarse-grained approach via argument component detection or a more fine-grained approach via the detection of argument properties.

For future research, we consider the following directions worth pursuing. First, while our classification results already show the potential of the annotations, we assume that models will benefit from more available data. Hence, an extension of the annotated dataset can be a fruitful task. Second, we currently work on adding a layer of span annotations to the corpus. We expect that expanding on the already existing full tweet annotations will facilitate this task. Third, we are also interested in investigating alternatives to traditional (monolingual) supervised machine learning approaches. For instance, cross-lingual AM (Eger et al., 2018) and few-shot learning (Wang et al., 2020) may be suitable ways to approach AM in scenarios where missing large annotated corpora are a bottleneck to progress in the field.

7. Acknowledgements

We thank Thomas Pham for assisting in data annotation.

8. Bibliographical References

Addawood, A. and Bashir, M. (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. Association for Computational Linguistics.

Aharoni, E., Polnarov, A., Lavee, T., Hershcovich, D., Levy, R., Rinott, R., Gutfreund, D., and Slonim, N. (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. Association for Computational Linguistics.

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., and Vollgraf, R. (2019). FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Al-Khatib, K., Wachsmuth, H., Hagen, M., Köhler, J., and Stein, B. (2016a). Cross-Domain Mining of Argumentative Text through Distant Supervision. In Kevin Knight, et al., editors, *12th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2016)*, pages 1395–1404. Association for Computational Linguistics.

Al-Khatib, K., Wachsmuth, H., Kiesel, J., Hagen, M., and Stein, B. (2016b). A News Editorial Corpus for Mining Argumentation Strategies. In Yuji Matsumoto et al., editors, *26th International Conference on Computational Linguistics (COLING 2016)*, pages 3433–3443. Association for Computational Linguistics.

Artstein, R. and Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Comput. Linguist.*, 34(4):555–596.

Bauwelinck, N. and Lefever, E. (2020). Annotating Topics, Stance, Argumentativeness and Claims in Dutch Social Media Comments: A Pilot Study. In *Proceedings of the 7th Workshop on Argument Mining*, pages 8–18, Online. Association for Computational Linguistics.

Beißwenger, M., Bartsch, S., Evert, S., and Würzner, K.-M. (2016). EmpiriST 2015: A Shared Task on the Automatic Linguistic Annotation of Computer-Mediated Communication and Web Corpora. In *Proceedings of the 10th Web as Corpus Workshop*, pages 44–56, Berlin. Association for Computational Linguistics.

Bhatti, M. M. A., Ahmad, A. S., and Park, J. (2021). Argument Mining on Twitter: A Case Study on the Planned Parenthood Debate. In *Proceedings of the 8th Workshop on Argument Mining*, pages 1–11, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bosc, T., Cabrio, E., and Villata, S. (2016a). 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. European Language Resources Association (ELRA).

Bosc, T., Cabrio, E., and Villata, S. (2016b). Tweeties squabbling: Positive and negative results in applying argument mining on social media. In *Computational Models of Argument - Proceedings of COMMA 2016*, volume 287 of *Frontiers in Artificial Intelligence and Applications*, pages 21–32, Potsdam, Germany. IOS Press.

Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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. Association for Computational Linguistics.

Dice, L. R. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26(3):297–302.

Dusmanu, M., Cabrio, E., and Villata, S. (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. Association for Computational Linguistics.

Eger, S., Daxenberger, J., Stab, C., and Gurevych, I. (2018). Cross-lingual Argumentation Mining: Machine Translation (and a bit of Projection) is All You Need! In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 831–844, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Goudas, T., Louizos, C., Petasis, G., and Karkaletsis, V. (2014). Argument Extraction from News, Blogs, and Social Media. In Aristidis Likas, et al., editors, *Artificial Intelligence: Methods and Applications*, pages 287–299, Cham. Springer International Publishing.

Lawrence, J. and Reed, C. (2020). Argument Mining: A Survey. *Computational Linguistics*, 45(4):765–818.

Levy, R., Bilu, Y., Hershcovich, D., Aharoni, E., and Slonim, N. (2014). Context Dependent Claim Detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Moens, M.-F., Boiy, E., Palau, R. M., and Reed, C. (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. Associ-

ation for Computing Machinery.

Park, J. and Cardie, C. (2014). Identifying Appropriate Support for Propositions in Online User Comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38, Baltimore, Maryland. Association for Computational Linguistics.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Proisl, T. and Uhrig, P. (2016). SoMaJo: State-of-the-art tokenization for German web and social media texts. In *Proceedings of the 10th Web as Corpus Workshop*, pages 57–62, Berlin. Association for Computational Linguistics.

Risch, J., Stoll, A., Wilms, L., and Wiegand, M. (2021). Overview of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 1–12, Duesseldorf, Germany. Association for Computational Linguistics.

Schaefer, R. and Stede, M. (2020). Annotation and Detection of Arguments in Tweets. In *Proceedings of the 7th Workshop on Argument Mining*, pages 53–58, Online. Association for Computational Linguistics.

Schaefer, R. and Stede, M. (2021). Argument Mining on Twitter: A survey. *it - Information Technology*, 63(1):45–58.

Stab, C. and Gurevych, I. (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. Association for Computational Linguistics.

Stede, M. and Schneider, J. (2018). *Argumentation Mining*, volume 40 of *Synthesis Lectures in Human Language Technology*. Morgan & Claypool.

Veltri, G. A. and Atanasova, D. (2017). Climate change on Twitter: Content, media ecology and information sharing behaviour. *Public Understanding of Science*, 26(6):721–737. PMID: 26612869.

Wang, Y., Yao, Q., Kwok, J. T., and Ni, L. M. (2020). Generalizing from a Few Examples: A Survey on Few-Shot Learning. *ACM Comput. Surv.*, 53(3):1–34.

Wilms, L., Heinbach, D., and Ziegele, M. (2021). Annotation guidelines for GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. Excerpt of an unpublished codebook of the DEDIS research group at Heinrich-Heine-University Düsseldorf (full version available on request).

Wührl, A. and Klinger, R. (2021). Claim Detection in Biomedical Twitter Posts. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 131–142, Online. Association for Computational Linguistics.