What to Do with an Airport? Mining Arguments in the German Online Participation Project Tempelhofer Feld

Matthias Liebeck[1] **Katharina Esau**[2] **Stefan Conrad**[1]

[1] Institute of Computer Science, Heinrich Heine University Düsseldorf, Germany
{liebeck,conrad}@cs.uni-duesseldorf.de

[2] Institute of Social Sciences, Heinrich Heine University Düsseldorf, Germany
katharina.esau@uni-duesseldorf.de

Abstract

This paper focuses on the automated extraction of argument components from user content in the German online participation project *Tempelhofer Feld*. We adapt existing argumentation models into a new model for decision-oriented online participation. Our model consists of three categories: major positions, claims, and premises. We create a new German corpus for argument mining by annotating our dataset with our model. Afterwards, we focus on the two classification tasks of identifying argumentative sentences and predicting argument components in sentences. We achieve macro-averaged F_1 measures of 69.77% and 68.5%, respectively.

1 Introduction

In the last few years in Germany, more and more cities are offering their citizens an internet-based way to participate in drafting ideas for urban planning or in local political issues. The administrations utilize websites to gather the opinions of their citizens and to include them in their decision making. For example, the German town Ludwigshafen has an elevated highway that is damaged and has to be demolished. Experts created four variants for a replacement and Ludwigshafen asked[1] its citizens to discuss them and to gather arguments for and against each variant, which were considered in the final political decision. Other cities such as Lörrach[2] tap into ideas of their citizens for a sustainable urban development and cities such as Darmstadt[3] and Bonn[4] also gather proposals in participatory budgetings. In general, these platforms are accompanied by offline events to inform residents and to allow for discussions with citizens who cannot or do not want to participate online. In the following, the term *online participation* refers to the involvement of citizens in relevant political or administrative decisions.

A participation process usually revolves around a specific subject area that is determined by the organizer. In a city, the administration might aim to collect ideas to improve a certain situation (e.g. how it can beautify a park). Aside from politics, companies or institutions can use online participation for policy drafting, for example, in universities (Escher et al., 2016).

By contrast, there are also platforms whose purpose is to report defects (e.g., such as a road in need of repair or a street lamp that needs replacing), which we do not regard further because they are only used for reporting issues and do not encourage discussions between citizens. In the scope of this paper, we focus only on the subset of online participation projects that aim to gather options for actions or decisions (e.g., "*We should build an opera.*" or "*Should we close the golf course or the soccer field?*"). Given a large number of suggestions and comments from citizens, we want to automatically identify options for actions and decisions, extract reasons for or against them (e.g., "*This would improve the cultural offerings of our city.*") and detect users' stances (e.g., "*I totally agree!*").

As far as we know, it is rather rare in a municipal administration that such participation processes can be attended to by full-time employees, because they have other responsibilities as well. If

[1] https://www.ludwigshafen-diskutiert.de

[2] https://gestalten.loerrach.de

[3] https://da-bei.darmstadt.de/discuss/Buergerhaushalt2014

[4] https://bonn-macht-mit.de

Proceedings of the 3rd Workshop on Argument Mining, pages 144–153,
Berlin, Germany, August 7-12, 2016.

a process is well received by the general public, it might attract hundreds of suggestions and thousands of comments. The manual analysis of this data is time consuming and could take months. Due to budgetary reasons, it might also not be possible to outsource the analysis. Is it therefore possible that an online participation process was a success and a large amount of text contributions has been created, but not all content can be taken into account. To avoid that huge amounts of text content become unprocessable, it is necessary to utilize automated techniques to ensure a contemporary processing. To the best of our knowledge, the automated extraction of argument components in the form of mining decision options and pro and contra arguments from German online participation projects in a political context is a research gap that we try to fill.

The remainder of the paper is structured as follows: Section 2 describes related work in argument mining. Section 3 explains our data, our annotation model and the annotation process. Our argument mining approach is described in section 4. We conclude and outline future work in section 5.

2 Related Work

Argumentation mining is an evolving research topic that deals with the automatic extraction of argument components from text. Most research focuses on English text, but there is also research for German (Houy et al., 2013; Habernal et al., 2014; Eckle-Kohler et al., 2015) and for the Greek language (Goudas et al., 2014).

Previous research spans a variety of domains, such as the legal domain (Palau and Moens, 2009; Houy et al., 2013), eRulemaking (Park and Cardie, 2014), student essays (Stab and Gurevych, 2014b), news (Eckle-Kohler et al., 2015), and web content (Goudas et al., 2014; Habernal and Gurevych, 2015). Most of the papers share common tasks, such as separating text into argumentative and non-argumentative parts, classifying argumentative text into argument components and identifying relations between them. Currently, there is no argument model that most researchers agree upon. The chosen argument model often depends on the tasks and the application domain. However, most of the recent research agrees that the two argument components *claim* and *premise* are usually part of the chosen argument models.

Most of the researched domains offer a high text quality. For instance, in the news domain, the text content is usually editorially reviewed before publishing. Since our text content is from the web, it partially lacks proper spelling or grammar and is sometimes difficult to understand. Nevertheless, it is important to develop methods for processing web content because everyone's opinion should be considered, especially in a political context.

Another characteristic of our application domain is the presence of discourse between different users. In an online participation platform, users often write comments that refer to other people's suggestions or justifications. This differs from other domains, such as newspaper articles and student essays, where text content is rather monologic.

To evaluate the performance of an argumentation mining system, datasets are humanly annotated (which results in a gold standard), for instance with argument components. More recent publications (Stab and Gurevych, 2014a; Park and Cardie, 2014; Habernal et al., 2014; Eckle-Kohler et al., 2015) report inter-annotator agreement values of how well multiple annotators agree on their annotations. Due to different available inter-annotator agreement measures and different annotation lengths (tokens, sentences or freely assignable spans), there is currently no standardized single measure for inter-annotator agreement in the argumentation mining community. As a result, we report multiple values to ensure better comparability in the future. A detailed overview of annotation studies can be found in (Habernal and Gurevych, 2016).

There has been previous research on automatically mining people's opinions in the context of political decisions named as *policy making* (Florou et al., 2013) and as *eRulemaking* (Park and Cardie, 2014; Park et al., 2015a), which relate to our application domain *online participation*.

Florou et al. (2013) web crawled Greek web pages and social media. The authors aim to extract arguments that are in support or in opposition of public policies. As a first step, they automatically classify text segments as argumentative or non-argumentative, although they do not describe what they consider as argumentative and they do not refer to argumentation theory. In our approach, we use text content from a specific platform (instead of crawling multiple sources); we

define three different argument components and their practical use; we relate to existing argumentation theory; and we further distinguish argument components in argumentative sentences.

Park and Cardie (2014) focus on English comments in the eRulemaking website *Regulation Room*. In their approach, they propose a model for eRulemaking that aims at verifiability by classifying propositions as *unverifiable*, *verifiable experiential*, and *verifiable non-experiential*. With their best feature set, they achieve a macro-averaged F_1 of 68.99% with a support vector machine. (Park et al., 2015b) discuss the results of conditional random fields as a machine learning approach. In our approach, we aim at identifying components and leave the issue of evaluability up to experts or to the wisdom of the crowd.

3 Data

This section discusses the data from the online participation project Tempelhofer Feld, presents our argumentation model, and describes our annotation process.

3.1 Background

The *Tempelhofer Feld*[5] project is an online participation project that focuses on the former airport *Berlin-Tempelhof (THF)* and its future use. Air traffic at the airport ceased in 2008. Until today, the 300 hectare area of the airport is mostly open space, which can be used for recreation.

In 2014, the *ThF-Gesetz (ThF law)* entered into force. It protects the large open space of the field, which is unique in Berlin, and limits structural changes, for example by prohibiting the construction of new buildings on the field.[6] The participation process was commissioned by Berlin's Senate Department for Urban Development and the Environment.

The project aims to collect ideas that improve the field for visitors while adhering to the ThF law, like settings up drinking fountains.

3.2 Discussion platform

The *Tempelhofer Feld* project uses the open-source policy drafting and discussion platform Adhocracy[7]. In Adhocracy, users can create proposals, which are text-based ideas or suggestions that contain a title and text content. To encourage discussions, users can comment on proposals and respond to previous comments, which results in a tree structured discussion per proposal. Adhocracy provides a voting system to upvote and downvote content. Therefore, users with limited time can follow *the wisdom of the crowd* by sorting proposals by their votes.

In the *Tempelhofer Feld* online participation process, users can register anonymously. Their voting behavior is public (it is possible to see which content was upvoted or downvoted by a specific user) and their text content is licensed under the Creative Commons License, which makes it attractive for academic use.

The official submission phase for proposals was from November 2014 until the end of March 2015. Afterwards, the proposals were condensed in offline events between May 2015 and July 2015. Until 2015-07-07, the users proposed 340 ideas and wrote 1389 comments. The comments vary in length. On average, they contain 3.56 sentences ($\sigma = 3.36$) and 58.7 tokens ($\sigma = 65.7$).

Each proposal is tagged with one out of seven categories. We excluded two categories because they mostly contain meta-discussions or serve as a "doesn't-fit-anywhere-else" category. This leaves the remaining five categories: *Bewirtschaftung* (cultivation), *Erinnerung* (memory), *Freizeit* (leisure), *Mitmachen* (participate), and *Natur* (nature).

The excluded categories are still important for the participation project, but, for the time being, we focus on proposals that contain ideas or suggestions that can potentially be realized. We do not judge whether it makes sense to realize the proposal or not. If someone wants to construct a roof over the whole area or wants to scatter blue pudding on the lawn, we leave it up to the other users to judge the proposal by voting and commenting on reasons for or against the realization, which we want to automatically extract.

We observed that the users occasionally did not use the platform correctly, by replying to a comment and referring to another previous comment.

It is worth mentioning that the participation process is not legally binding and that the most upvoted proposals do not become realized automatically. Although the participation process is encouraged by the politicians, the final decision which proposals will be realized is still up to them.

[5] https://tempelhofer-feld.berlin.de

[6] There are a few exceptions, like lighting, sanitary facilities, seating, and waste bins.

[7] https://github.com/liqd/adhocracy

3.3 Argumentation Model

We have a practical point of view on text content in political online participation processes: To allow politicians to include the opinions expressed in the platform into their decision making, we need to extract three different components: (i) what do people want to be built or decided upon, (ii) how do people argue for and against these ideas, and (iii) how many people in the discussion say that they agree or disagree with them.

First, we tried to apply existing argumentation models that are commonly used in argument mining to our dataset, namely Toulmin's model (Toulmin, 1958) and the claim-premise model (based on (Freeman, 1991)). We quickly realized that attacks on logical conclusions are rather rare, that users frequently express their wishes and participate by providing reasons for and against other suggestions, and that we have to consider this behavior in the choice of an argumentation model.

Toulmin differentiates between six argument components: *claim*, *ground / data*, *warrant*, *backing*, *qualifier* and *rebuttal*. The model revolves around the claim, the statement of the argument which has to be proven or, in Toulmin's words, "*whose merits we are seeking to establish*" (Toulmin, 2003, p. 90). Grounds are the data that support the claim and serve "*as a foundation for the claim*" (Toulmin, 2003, p. 90). A ground is connected to the claim by a warrant, which justifies why the ground supports the claim. A warrant can be supported by a backing which establishes "*authority*" (Toulmin, 2003, p. 96) as to why the warrant is to be accepted. A qualifier specifies the degree of certainty or the "*degree of force*" (Toulmin, 2003, p. 93) of the claim, in respect of the ground. Rebuttals are conditions which "*might be capable of defeating*" (Toulmin, 2003, p. 94) the claim. With Toulmin's model, we come to the same conclusion as Habernal et al. (2014) that it is difficult to apply the model to online-generated discussions, especially when the users argue on a level where most of Toulmin's categories can only be applied very rarely.

The commonly used claim-premise model (Palau and Moens, 2009; Peldszus and Stede, 2013; Stab and Gurevych, 2014a; Eckle-Kohler et al., 2015) consists of the two components claim and premise. A *claim* "*is the central component of an argument*" (Stab and Gurevych, 2014a), whose merit is to be established. *Premises* are reasons that either support or attack a claim. According to Stab and Gurevych (2014a), a claim "*should not be accepted by readers without additional support.*" Palau and Moens (2009) describe a claim as "*an idea which is either true or false*" and Stab and Gurevych (2014a) as a "*controversial statement that is either true or false.*"

We share the opinion of Habernal et al. (2014) that there is no one-size-fits-all argumentation theory for web data and follow the recommendation that the argumentation model should be chosen for the task at hand. In our participation project, we are primarily interested in mining suggestions. This differs from the common focus on mining claims as the central component, because the definition of a claim stated above does not apply to our dataset: suggestions cannot be classified as true or false and they can be accepted without additional support, although justifications are commonly provided by the users.

We adapted the claim-premise family and its modification for persuasive essays in Stab and Gurevych (2014a) to a three-part model for modeling arguments in online participation processes: (i) major positions, (ii) claims, and (iii) premises

Major positions are options for actions or decisions that occur in the discussion (e.g., "*We should build a playground with a sandbox.*" or "*The opening hours of the museum must be at least two hours longer.*"). They are most often someone's vision of something new or of a policy change. If another user suggests a modified version by changing some details, the new suggestion is a new major position (e.g. "*We should build a playground with swings.*"). In our practical view, major positions are unique suggestions from citizens that politicians can decide on.

A **claim** is a pro or contra stance towards a major position (e.g. "*Yes, we should definitely do that!*"). In our model, claims are text passages in which users express their personal positionings (e.g., "*I dislike your suggestion.*"). For a politician, the text content of a claim in our definition does not serve as a basis for decision making because claims do not contain justifications upon which decisions can be backed up. The purpose behind mining these claims is a conversion into two numbers that indicates how many citizens are for or against a suggestion.

The term **premise** is defined as a reason that attacks or supports a major position, a claim or

another premise. Premises are used to make an argumentation comprehensible for others, by reasoning why a suggestion or a decision should be realized or why it should be avoided (e.g. "*This would allow us to save money.*"). We use the term premise in the same way as the claim-premise model and as Toulmin with *grounds*.

We do not evaluate if a reason is valid. We only determine the user's intent: If an annotator perceives that a user is providing a reason, we annotate it as such. Otherwise, we would have to evaluate each statement on a semantic level. For example, if a user argues that a suggestion violates a building law, the annotators would need to check this statement. A verification of all reasons for correctness would require too much expertise from annotators or a very large knowledge database. In our application domain, we leave the evaluation up to human experts who advise politicians.

Our argumentation model is illustrated in Figure 1.

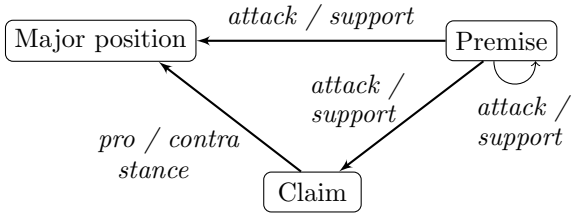

Figure 1: Our argumentation model for political online participation

If a sentence contains only one argument component, we annotate the whole sentence. If there is more than one argument component in a sentence, we annotate the different components separately, like in the following example: [*claim*: We don't need a new playground] [*premise*: because we already have one.]

Depending on the writing style of a user, a thought or idea might be expressed in more than one sentence. In such a case, we combine all successive sentences of the same thought into a group: [*major position*: The city should build a public bath. It should contain a 50 meter pool and be flooded with daylight.] The boundaries of these groups are based on the content and, therefore, are subjective. Thus, in our evaluation, we first focus on a sentence-based classification of argument components and consider the identification of the group boundaries as future work. Freeman (1991, p. 106) uses the term *linked* for premises that consist of multiple statements, each of which does not separately constitute a support, but together produce "*a relevant reason*".

Major positions are very similar to the concept of *policy claims* which "*advocate a course of action*" and are about "*deciding what to do*" (Schiappa and Nordin, 2013, p. 101).

3.4 Annotation

We developed annotation guidelines and refined them over the course of multiple iterations. Our dataset was annotated by three annotators of which two are authors of this publication. The third annotator was instructed after the annotation guidelines were developed.

OpenNLP[8] was used to split each proposal and comment into individual sentences. Errors were manually corrected. We also removed embedded images that occur sporadically because we focus on text content. Afterwards, we used the *brat rapid annotation tool* (Stenetorp et al., 2012) for the annotation of the dataset. The text content also contains non-argumentative sentences which we did not annotate. These include salutations, valedictions, meta-discussions (for instance, comments about the participation process), and comprehension questions.

In our annotation process, we further divide claims into *pro claims* and *contra claims* by classifying the most dominant positioning, based on the content and the wording in order to report a simplified "level of agreement / disagreement" in preparation for a future user behavior study. More observations of our annotation process are detailed in section 6.

3.4.1 Inter-Annotator Agreement

Before annotating the data set, we took a subset of 8 proposals with 74 comments to measure the inter-annotator-agreement, consisting of 261 sentences and 4.1k tokens. The subset was randomly drawn from 67 proposals that have between 5 and 40 comments.

As in recent research (Stab and Gurevych, 2014a; Park and Cardie, 2014; Habernal et al., 2014; Eckle-Kohler et al., 2015), we also report inter-annotator agreement (IAA) values to quantify the consensus of our annotators and to make our annotation study more comparable. As there is

[8] https://opennlp.apache.org

	$A_{o,t}$	κ_t	α_u
all	76.4	62.6	78.0
major positions	89.3	71.9	79.8
claims pro	96.3	66.1	59.0
claims contra	95.6	52.3	57.2
premises	80.9	61.5	80.1
AU / non-AU	90.7	49.1	92.4

Table 1: Inter-annotator agreement scores in percentages: $A_{o,t}$ token-based observed agreement, κ_t token-based Fleiss' kappa, and α_u Krippendorff's unitized alpha

currently no standardized single measure in the argumentation mining community, we report multiple IAA values. We use *DKPro Agreement* (Meyer et al., 2014) to report our inter-annotator agreement values. Table 1 summarizes our IAA values for three scenarios: (i) joint measures over all categories, (ii) category-specific values, and (iii) argumentative vs. non-argumentative units.

Since we asked the annotators to assign *labels* to *freely assignable spans*, we use *Krippendorff's unitized alpha* α_u (Krippendorff, 2004). We have to keep in mind that several comments only contain one sentence and are, therefore, much easier to annotate. An average over IAA values from all comments would be biased. Hence, we follow the proposed approach in (Habernal and Gurevych, 2016) to concatenate all text content into a single document and measure a single Krippendorff's α_u value instead of averaging α_u for each document.

We also report the token-based *observed agreement* $A_{o,t}$ and the token-based *Fleiss' kappa* κ_t (Fleiss, 1971). The token-based distribution of the annotions of all three annotators is as follows: 1278 non-argumentative tokens and 11220 argumentative tokens (3214 major positions, 730 claims pro, 583 claims contra, 6693 premises)

We do not report a sentence-based inter-annotator agreement because more than one annotation per sentence is possible (e.g., a claim followed by a premise in a subordinate clause) and the IAA measures are for single-label annotation only.

The measures of $\alpha_u = 0.924$ for argumentative versus non-argumentative spans and the joint measure for all categories of $\alpha_u = 0.78$ indicate a reliable agreement between our three annotators. Therefore, we should be able to provide good annotations for automated classification tasks.

3.4.2 Corpus

For our corpus, we randomly drew 72 proposals that each contain at least one major position. These proposals were commented with 575 comments. In total, our annotated dataset consists of 2433 sentences and 40177 tokens. We annotated 2170 argumentative spans. They comprise 548 major positions, 378 claims (282 pro claims and 96 contra claims), and 1244 premises. Our annotated corpus consists of 4646 (11.6%) non-argumentative and 35531 (88.4%) argumentative tokens. This indicates that the text content is highly argumentative. Exactly 88 (3.6%) of the sentences were annotated with more than one argument component.

We plan to release our dataset along with our annotations under an open-source license to allow reproducibility.

4 Evaluation

This section discusses our initial approach to automatically identify argumentative sentences and to classify argument components.

4.1 Preprocessing

First, we tokenize all sentences in our dataset with *OpenNLP* and use *Mate Tools* (Björkelund et al., 2010) for POS-tagging and dependency parsing.

4.2 Features

For our classification problems, we evaluate different features and their combinations. They can be divided into three groups: (i) n-grams, (ii) grammatical distributions, and (iii) structural features. N-grams are an obvious choice to capture the text content because several words are used repeatedly in different argument components, like "*agree*" or "*disagree*" in the case of claims. We use unigrams and bigrams as binary features.

Grammatical Distributions Based on our observations, we identified that users use different tenses and sentences structures for our three categories. For instance, claims are often stated in the present tence (e.g., "*I agree!*"). Therefore, we use an L_2-normalized POS-Tag distribution of the STTS tags (Schiller et al., 1999) and an L_2-normalized distribution of the dependencies in the TIGER annotation scheme (Albert et al., 2003).

Structural Features We also capture multiple structural features: token count, percent-

Feature Set	AU / non-AU			Argument Components		
	SVM	RF	k-NN	SVM	RF	k-NN
Unigram	65.99	68.13	61.00	64.40	59.41	40.30
Unigram, lowercased	66.69	64.53	62.26	65.32	53.35	38.25
Bigram	41.79	50.48	16.25	46.62	50.42	11.51
Grammatical	55.88	52.24	48.52	59.54	47.89	46.81
Unigram + Grammatical	**69.77**	58.39	64.87	**68.50**	57.13	35.90
Unigram + Grammatical + Structural	67.50	61.14	54.07	65.99	59.46	47.27

Table 2: Macro-averaged F_1 scores for the two classification problems: (i) classifying sentences as argumentative and non-argumentative, (ii) classifying sentences as major positions, claims, and premises.

age of comma tokens in the sentence, percentage of dot tokens in the sentence, and the last token of a sentence as an one-hot encoding ('.', '!', '?', 'OTHER'). Furthermore, we use the index of the sentence since we have noticed that users often start their comment with a pro or contra claim. Moreover, we use the number of links in a sentence as a feature.

4.3 Results

We report results for two classification problems. Subtask A is the classification of sentences as argumentative or non-argumentative and in subtask B we automatically classify argument components in sentences with exactly one annotated argument component. Macro-averaged F_1 was chosen as evaluation metric. For each subtask, we randomly split the respective annotated sentences into a 80% training set and 20% test set.

Different feature combinations were evaluated with three classifiers: Support vector machine (SVM) with an RBF kernel, random forest (RF), and k-nearest neighbor (k-NN). We use *scikit-learn* (Pedregosa et al., 2011) as machine learning library. The required parameters for our classifiers (SVM: penalty term C and γ for the kernel function; random forest: number of trees, maximal depth of the trees, and multiple parameters regarding splits; k-NN: number of neighbors k and weight functions) were estimated by a grid search on a 10-fold cross-validation on the training set.

The results of both subtasks are listed in Table 2. k-NN almost always achieved the worst results in comparison with the two classifiers. The results of bigrams as features are worse than the results of unigrams. Lowercasing words has different effects, depending on the classifier: The results of unigrams improve for SVMs but decline for random forests and k-NN. The addition of the structural features also had different effects on the classifiers, depending on the subtask. Additionally, we experimented with lemmatized words by Mate Tools (combined with *IWNLP* (Liebeck and Conrad, 2015)) but our results were slightly lower. In our future work, we will work on better ways to incorporate lemmatization into our classification tasks.

4.3.1 Subtask A

For identifying argumentative sentences, the best result of 69.77% was achieved by a support vector machine with unigrams and grammatical features. It is interesting to see that *unigrams* work better with the random forest classifier than with an SVM, but, with the additional *grammatical* features, the SVM outperforms the random forest. The training set for subtask A contains 1667 argumentative and 280 non-argumentative sentences, whereas the test set comprises 411 argumentative and 75 non-argumentative sentences.

4.3.2 Subtask B

For the classification of argument components, we do not further differentiate between pro and contra claims because both of them occur rarer than major positions and premises. Therefore, we have grouped pro and contra claims. The training set for subtask B contains 1592 sentences (951 premises, 399 major positions, and 242 claims), whereas the test set comprises 398 sentences (219 premises, 110 major positions, and 69 claims).

The best result for subtask B with a macro-averaged F_1 score of 68.5% was again achieved by a support vector machine as classifier with unigrams and grammatical features. In subtask B, the gap between the results of the k-NN classifier and the results of the two classifiers is much larger than in subtask A.

		Predicted			
		MP	C	P	$\sum$
Actual	MP	63	4	43	110
	C	9	48	12	69
	P	27	20	172	219
	$\sum$	99	72	227	398

Table 3: Confusion matrix for our best result of identifying argument components with a support vector machine and "*unigram + grammatical*" as features

In order to better understand our results, we report the confusion matrix for the best classifier in Table 3. The confusion matrix shows that the classification of premises works well and that major positions are often misclassified as premises. In our future work, we will try to find better semantic features to differentiate major positions from premises.

We initially tried to solve subtask B as a four class problem but our features do not allow for a good distinction between pro and contra claims with our small training size for claims yet. In our future work, we will treat their distinction as a further classification task and will integrate more polarity features.

5 Conclusion and Future Work

In this paper we have presented a new corpus for German argumentation mining and a modified argumentation scheme for online participation. We described the background of our data set, our annotation process, and our automated classification approach for the two classification tasks of identifying argumentative sentences and identifying argument components. We evaluated different feature combinations and multiple classifiers. Our initial results for argument mining in the field of German online participation are promising. The best results of 69.77% in subtask A and 68.5% in subtask B were both achieved by a support vector machine with unigrams and grammatical features.

While working with our dataset, we realized that citizens argue not only with rational reasons and that they are not always objective. They often express their positive and negative emotions and use humor to convince other participants or just to avoid conflicts. This makes an automatic approach more difficult.

In our future work, we want to experiment with additional features to further increase our classification results. We will identify specific emotions in the argumentation among citizens. We will try to find humor as a predictor for enjoyment and sociability.

So far, we have only worked on a sentence level. We would like to automatically detect tokens that form a group, based on the content. For this, we could use the token-based BIO scheme used in Goudas et al. (2014) and Habernal and Gurevych (2016), which divides tokens into beginning (B), inner (I), and other (O) tokens of an argument component. This would also allow us to find more than one argument component in a sentence.

Furthermore, we will work on the distinction of claims into pro and contra claims. Additionally, we aim to identify more freely available corpora for online participation to which we can apply our model for a comparative study.

6 Observations

Background knowledge Some proposals and comments require background knowledge in order to fully comprehend them. For an automated approach, this is much more difficult, especially if existing buildings on the field or city districts are referred to by name.

Edge annotation We chose not to annotate outgoing edges in our corpus. In a single label approach, ambiguity might occur because a premise might support one claim and attack another one. We tried an approach with multiple outgoing edges but it became very difficult to evaluate every possible edge in discussions with more than 30 comments and multiple major positions. In order to avoid an incomplete edge annotation, we completely omitted the annotation of edges for the time being.

Contextual differentiation During the annotation, we noticed some situations where it became difficult to decide which argument component is the best fit. For instance, "*Vertical vegetable gardens are an enrichment for our perception.*" contains a slight positioning, but in the context of the comment, the sentence was used as a reason and, therefore, annotated as a premise.

Acknowledgments

This work was funded by the PhD program *Online Participation*, supported by the North Rhine-

Westphalian funding scheme *Fortschrittskollegs*. The authors want to thank the anonymous reviewers for their suggestions and comments.

References

Stefanie Albert, Jan Anderssen, Regine Bader, Stephanie Becker, Tobias Bracht, Sabine Brants, Thorsten Brants, Vera Demberg, Stefanie Dipper, Peter Eisenberg, Silvia Hansen, Hagen Hirschmann, Juliane Janitzek, Carolin Kirstein, Robert Langner, Lukas Michelbacher, Oliver Plaehn, Cordula Preis, Marcus Pussel, Marco Rower, Bettina Schrader, Anne Schwartz, Smith George, and Hans Uszkoreit. 2003. TIGER Annotationsschema. Technical report, Universität Potsdam, Universität Saarbrücken, Universität Stuttgart.

Anders Björkelund, Bernd Bohnet, Love Hafdell, and Pierre Nugues. 2010. A High-Performance Syntactic and Semantic Dependency Parser. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, COLING '10, pages 33–36. Association for Computational Linguistics.

Judith Eckle-Kohler, Roland Kluge, and Iryna Gurevych. 2015. On the Role of Discourse Markers for Discriminating Claims and Premises in Argumentative Discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, pages 2236–2242.

Tobias Escher, Dennis Friess, Katharina Esau, Jost Sieweke, Ulf Tranow, Simon Dischner, Philipp Hagemeister, and Martin Mauve. 2016. Online Deliberation in Academia: Evaluating the Quality and Legitimacy of Co-Operatively Developed University Regulations. *Policy & Internet*, (in press).

Joseph L. Fleiss. 1971. Measuring Nominal Scale Agreement Among Many Raters. *Psychological Bulletin*, 76(5):378 – 382.

Eirini Florou, Stasinos Konstantopoulos, Antonis Koukourikos, and Pythagoras Karampiperis. 2013. Argument extraction for supporting public policy formulation. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 49–54. Association for Computational Linguistics.

James Freeman. 1991. *Dialectics and the Macrostructure of Arguments: A Theory of Argument Structure*, volume 10 of *Pragmatics and Discourse Analysis Series*. de Gruyter.

Theodosios Goudas, Christos Louizos, Georgios Petasis, and Vangelis Karkaletsis. 2014. Argument Extraction from News, Blogs, and Social Media. In *Artificial Intelligence: Methods and Applications*, pages 287–299.

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 (EMNLP)*, pages 2127–2137. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2016. Argumentation Mining in User-Generated Web Discourse. *Computational Linguistics*, (in press).

Ivan Habernal, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Argumentation Mining on the Web from Information Seeking Perspective. In *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*, pages 26–39. CEUR-WS.

Constantin Houy, Tim Niesen, Peter Fettke, and Peter Loos. 2013. Towards Automated Identification and Analysis of Argumentation Structures in the Decision Corpus of the German Federal Constitutional Court. In *7th IEEE International Conference on Digital Ecosystems and Technologies (IEEE DEST)*. IEEE Computer Society.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, second edition.

Matthias Liebeck and Stefan Conrad. 2015. IWNLP: Inverse Wiktionary for Natural Language Processing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 414–418. Association for Computational Linguistics.

Christian Meyer, Margot Mieskes, Christian Stab, and Iryna Gurevych. 2014. DKPro Agreement: An Open-Source Java Library for Measuring Inter-Rater Agreement. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 105–109.

Raquel 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, ICAIL '09*, pages 98–107. ACM.

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. Association for Computational Linguistics.

Joonsuk Park, Cheryl Blake, and Claire Cardie. 2015a. Toward Machine-assisted Participation in eRulemaking: An Argumentation Model of Evaluability. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, ICAIL '15, pages 206–210. ACM.

Joonsuk Park, Arzoo Katiyar, and Bishan Yang. 2015b. Conditional Random Fields for Identifying Appropriate Types of Support for Propositions in Online User Comments. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 39–44. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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. Association for Computational Linguistics.

Edward Schiappa and John Nordin. 2013. *Argumentation: Keeping Faith with Reason*. Pearson Education.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. Guidelines für das Tagging deutscher Textcorpora mit STTS (kleines und großes Tagset). Technical report, Universität Stuttgart, Universität Tübingen.

Christian Stab and Iryna Gurevych. 2014a. Annotating Argument Components and Relations in Persuasive Essays. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, pages 1501–1510.

Christian Stab and Iryna Gurevych. 2014b. Identifying Argumentative Discourse Structures in Persuasive Essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 46–56.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 102–107. Association for Computational Linguistics.

Stephen Toulmin. 1958. *The Uses of Argument*. Cambridge University Press.

Stephen Toulmin. 2003. *The Uses of Argument, Updated Edition*. Cambridge University Press.