Argumentation Mining in User-Generated Web Discourse

Ivan Habernal*
German Institute for Educational Research and Technische Universität Darmstadt

Iryna Gurevych**
Technische Universität Darmstadt and German Institute for Educational Research

The goal of argumentation mining, an evolving research field in computational linguistics, is to design methods capable of analyzing people's argumentation. In this article, we go beyond the state of the art in several ways. (i) We deal with actual Web data and take up the challenges given by the variety of registers, multiple domains, and unrestricted noisy user-generated Web discourse. (ii) We bridge the gap between normative argumentation theories and argumentation phenomena encountered in actual data by adapting an argumentation model tested in an extensive annotation study. (iii) We create a new gold standard corpus (90k tokens in 340 documents) and experiment with several machine learning methods to identify argument components. We offer the data, source codes, and annotation guidelines to the community under free licenses. Our findings show that argumentation mining in user-generated Web discourse is a feasible but challenging task.

1. Introduction

The art of **argumentation** has been studied since the early work of Aristotle, dating back to the 4th century BCE (Aristotle and Kennedy [translator] 1991). It has been exhaustively examined from different perspectives, such as philosophy, psychology,

* Ubiquitous Knowledge Processing Lab (UKP-DIPF), German Institute for Educational Research, Schloßstraße 29, D-60486 Frankfurt am Main, Germany and Technische Universität Darmstadt, Ubiquitous Knowledge Processing (UKP) Lab, TU Darmstadt - FB 20a Hochschulstrasse 10, D-64289 Darmstadt, Germany. E-mail: habernal@ukp.informatik.tu-darmstadt.de

** Technische Universität Darmstadt, Ubiquitous Knowledge Processing (UKP) Lab, TU Darmstadt - FB 20a Hochschulstrasse 10, D-64289 Darmstadt, Germany and Ubiquitous Knowledge Processing Lab (UKP-DIPF), German Institute for Educational Research, Schloßstraße 29, D-60486 Frankfurt am Main, Germany.

Submission received: 2 April 2015; revised version received: 20 April 2016; accepted for publication: 14 June 2016.

doi:10.1162/COLI_a_00276

communication studies, cognitive science, formal and informal logic, linguistics, computer science, educational research, and many others. In a recent and critically well-acclaimed study, Mercier and Sperber (2011) even claim that argumentation is what drives humans to perform reasoning. From the pragmatic perspective, argumentation can be seen as a *verbal activity oriented towards the realization of a goal* (Micheli 2011) or more in detail as *a verbal, social, and rational activity aimed at convincing a reasonable critic of the acceptability of a standpoint by putting forward a constellation of one or more propositions to justify this standpoint* (van Eemeren, Grootendorst, and Snoeck Henkemans 2002).

Analyzing argumentation from the computational linguistics point of view has very recently led to a new field called **argumentation mining** (Green et al. 2014). Despite the lack of an exact definition, researchers within this field usually focus on analyzing discourse on the pragmatics level and applying a certain argumentation theory to model and analyze textual data[1] at hand.

Our motivation for argumentation mining stems from a practical *information seeking* perspective from the user-generated content on the Web. For example, when users search for information in user-generated Web content to facilitate their personal decision-making related to controversial topics, they lack tools to overcome the current information overload. One particular use-case example dealing with a forum post discussing *private versus public schools* is shown in Figure 1. Here, the lengthy text on the left-hand side is transformed into an argument gist on the right-hand side by (i) analyzing argument components and (ii) summarizing their content. Figure 2 shows another use-case example, in which users search for reasons that underpin certain standpoints in a given controversy (which is *homeschooling* in this case). In general, the output of automatic argument analysis performed on the large scale in Web data can provide users with analyzed arguments to a given topic of interest, find the evidence for the given controversial standpoint, or help to reveal flaws in argumentation of others.

Satisfying these information needs cannot be directly tackled by current methods (e.g., opinion mining, question answering,[2] or summarization[3]), and requires novel approaches within the argumentation mining field. Although user-generated Web content has already been considered in argumentation mining, many limitations and research gaps can be identified in the existing research. First, the scope of the current approaches is restricted to a particular domain or register—for example, hotel reviews (Wachsmuth et al. 2014), Tweets related to local riot events (Llewellyn et al. 2014), student essays (Stab and Gurevych 2014a), airline passenger rights and consumer protection (Park and Cardie 2014), or renewable energy sources (Goudas et al. 2014). Second, not all related works are tightly connected to argumentation theories, resulting in a gap between the substantial research in argumentation itself and its adaptation in natural language processing (NLP) applications. Third, as an emerging research area, argumentation mining still suffers from a lack of labeled corpora, which is crucial for designing, training, and evaluating the algorithms. Although some works have dealt with creating new data sets, the reliability (in terms of inter-annotator agreement) of the

1 Despite few recent multi-modal approaches to process argumentation and persuasion (e.g., Brilman and Scherer 2015), the main mode of argumentation mining is natural language text.

2 These research fields are still related and complementary to argumentation mining. For example, personal decision-making queries (such as *Should I homeschool my children?*) might be tackled by research exploiting social question-answering sites.

3 The role of argumentation moves in summarizing scientific articles was examined by Teufel and Moens (2002).

Original text

The public schooling system is not as bad as some may think. Some mentioned that those who are educated in the public schools are less educated, well I actually think it would be in the reverse. Student who study in the private sector actually pay a fair amount of fees to do so and I believe that the students actually get let off for a lot more than anyone would in a public school. And its all because of the money. In a private school, a student being expelled or suspended is not just one student out the door, its the rest of that students schooling life fees gone. Whereas in a public school, its just the student gone. I have always gone to public schools and when I finished I got into University. I do not feel disadvantaged at all.

Extracted argument gist

I claim that public schools are good **because** students in private schools are only source of money. **I can back-up my argument:** I have always gone to public schools and when I finished I got into University. I do not feel disadvantaged at all. **On the other hand** some mentioned that those who are educated in the public schools are less educated.

Figure 1
Motivation example 1: Extracting argument gist by means of analyzing the argument structure and summarizing the argument components. The **bold** phrases are generated automatically and invoke the component function; Madnani et al. (2012) refers to these organizational elements as *shells*. *Doc#4733, forum post, public-private schools*.

Reasons for homeschooling

- Schools provide a totally unstimulating environment.
- Lesson plans (and the national curriculum) are the death of real education.
- Evidence including our own suggests strongly that this kind of education prepares children to enter further and higher education, or the workforce - and offers them the freedom to learn in the ways that suit them best.
- We teach our children how to learn, not merely how to pass tests.

Reasons against homeschooling

- Keeping your kids away from knowledge that you don't like is a moral crime.
- Religious zealotry is no excuse for raising a kid devoid of a proper education.
- Consciously depriving a child of an adequate education solely because "father knows best," or thinks he does, is tantamount to child abuse.

Figure 2
Motivation example 2: Extracting evidence for a certain standpoint with respect to a given controversial topic. All statements are taken from the corpus introduced in this article.

annotated resources is often unknown (Feng and Hirst 2011; Mochales and Moens 2011; Walton 2012; Villalba and Saint-Dizier 2012; Florou et al. 2013).

Annotating and automatically analyzing arguments in unconstrained user-generated Web discourse represent challenging tasks. So far, the research in argumentation mining "has been conducted on domains like news articles, parliamentary records and legal documents, where the documents contain well-formed explicit arguments, i.e., propositions with supporting reasons and evidence present in the text" (Park and Cardie 2014, page 29). Boltužić and Šnajder (2014, page 50) point out that "unlike in debates or other more formal argumentation sources, the arguments provided by the users, if any, are less formal, ambiguous, vague, implicit, or often simply poorly worded." Another challenge stems from the different nature of argumentation theories and computational linguistics. Whereas computational linguistics is mainly descriptive, the empirical research that is carried out in argumentation theories does not constitute a test of the theoretical model that is favored, because the model of argumentation is a *normative* instrument for assessing the argumentation (van Eemeren et al. 2014, page 11).

So far, no fully fledged descriptive argumentation theory based on empirical research has been developed, thus the feasibility of adapting argumentation models to the Web discourse represents an open issue.

These challenges can be formulated into the following research questions:

- Can we adapt models from argumentation theories that have been usually lacking empirical evidence on large real-world corpora for modeling argumentation in user-generated Web content?
- What are the desired properties of the argumentation model and is there a trade-off between model complexity and annotation reliability?
- What phenomena are typical of argumentation on the Web, how should we approach their modeling, and what challenges do they pose?
- What is the impact of different controversial topics and are there differences in argumentation between various registers?
- What computational approaches can be used to analyze arguments on the Web?

In this article, we push the boundaries of the argumentation mining field by focusing on several novel aspects. We tackle these research questions as well as the previously discussed challenges and issues. First, we target user-generated Web discourse from several domains across various registers to examine how argumentation is communicated in different contexts. Second, we bridge the gap between argumentation theories and argumentation mining by selecting the argumentation model based on research into argumentation theories and related fields in communication studies or psychology. In particular, we adapt normative models from argumentation theory to perform empirical research in NLP and support our application of argumentation theories with an in-depth reliability study. Finally, we use state-of-the-art NLP techniques in order to build robust computational models for analyzing arguments that are capable of dealing with a variety of genres on the Web.[4]

1.1 Our Contributions

We create a **new corpus** which is, to the best of our knowledge, the largest corpus that has been annotated within the argumentation mining field to date. We choose several target domains from educational controversies, such as *homeschooling*, *single-sex education*, and *mainstreaming*.[5] A novel aspect of the corpus is its coverage of different registers of **user-generated Web content**, such as comments on articles, discussion forum posts, blog posts, as well as professional newswire articles.

Because the data come from a variety of sources and no assumptions about their actual content with respect to argumentation can be drawn, we conduct two extensive

4 We used the data set and core methods from this article in our subsequent publication (Habernal and Gurevych 2015). The main difference is that this article focuses mainly on corpus annotation, analysis, and argumentation on Web data in general, whereas in Habernal and Gurevych (2015) we explored whether methods for recognizing argument components benefit from using semi-supervised features obtained from noisy debate portals.

5 Controversial educational topics attract a wide range of participants, such as parents, journalists, education experts, policy makers, and students, which contributes to the linguistic breadth of the discourse.

annotation studies. In the first study, we tackle the problem of relatively high "noise" in the retrieved data. In particular, not all of the documents are related to the given topics in a way that makes them candidates for further deep analysis of argumentation (this study results into 990 annotated documents). In the second study, we discuss the selection of an appropriate argumentation model based on evidence in argumentation research and propose a model that is suitable for analyzing micro-level argumentation in user-generated Web content. Using this model, we annotate 340 documents (approx. 90,000 tokens), reaching substantial inter-annotator agreement. We provide a hand analysis of all the phenomena typical to argumentation that are prevalent in our data. These findings may also serve as empirical evidence to issues that are at the forefront of current argumentation research.

From the computational perspective, we experiment on the annotated data using various machine learning methods in order to extract argument structure from documents. We propose several novel **feature sets** and identify configurations that run best in in-domain and cross-domain scenarios. To foster research in the community, we provide the annotated data as well as all the experimental software under free license.[6]

The rest of the article is structured as follows. First, we provide an essential background in argumentation theory in Section 2. Section 3 surveys related work in several areas. Then we introduce the data set and two annotation studies in Section 4. Section 5 presents our experimental work and discusses the results and errors, and Section 6 concludes this article.

2. Theoretical Background

Let us first present some definitions of the term **argumentation** itself. Ketcham (1917, page 3) defines argumentation as *the art of persuading others to think or act in a definite way. It includes all writing and speaking which is persuasive in form.* According to MacEwan (1898), *argumentation is the process of proving or disproving a proposition. Its purpose is to induce a new belief, to establish truth or combat error in the mind of another.* Freeley and Steinberg (2008, page 2) narrow the scope of argumentation to *reason giving in communicative situations by people whose purpose is the justification of acts, beliefs, attitudes, and values.* Although these definitions vary, the purpose of argumentation remains the same—to persuade others.

We would like to stress that our perception of argumentation goes beyond somehow limited *giving reasons* (Freeley and Steinberg 2008; Damer 2013). Rather, we see the goal of argumentation as to persuade (Ketcham 1917; Mercier and Sperber 2011; Nettel and Roque 2011). **Persuasion** can be defined as *a successful intentional effort at influencing another's mental state through communication in a circumstance in which the persuadee has some measure of freedom* (O'Keefe 2002, page 5); although, as O'Keefe (2011) points out, there is no correct or universally endorsed definition of either "persuasion" or "argumentation." However, broader understanding of argumentation as a *means* of persuasion allows us to take into account not only reasoned discourse but also non-reasoned mechanisms of influence, such as emotional appeals (Blair 2011).

Having an **argument** as a product within the argumentation process, we should now define it. One typical definition is that *an argument is a claim supported by reasons* (Schiappa and Nordin 2013, page 6). The term **claim** has been used since the 1950's, introduced by Toulmin (1958)—and in argumentation theory it is a synonym for

6 https://www.ukp.tu-darmstadt.de/data/argumentation-mining/.

standpoint or *point of view*. It refers to what is at issue in the sense of what is being argued about. The presence of a standpoint is thus crucial for argumentation analysis. However, the claim as well as other parts of the argument might be implicit; this is known as **enthymematic argumentation**, which is rather usual in ordinary argumentative discourse (Amossy 2009).

One fundamental problem with the definition and formal description of arguments and argumentation is that there is no agreement even among argumentation theorists. As van Eemeren et al. (2014, page 29) admit in their very recent and exhaustive survey of the field, "as yet, there is no unitary theory of argumentation that encompasses the logical, dialectical, and rhetorical dimensions of argumentation and is universally accepted. The current state of the art in argumentation theory is characterized by the coexistence of a variety of theoretical perspectives and approaches, which differ considerably from each other in conceptualization, scope, and theoretical refinement."

2.1 Argumentation Models

Despite the missing consensus on the ultimate argumentation theory, various argumentation models have been proposed that capture argumentation on different levels. Argumentation models abstract from the language level to a concept level that stresses the links between the different components of an argument or how arguments relate to each other (Prakken and Vreeswijk 2002). Bentahar, Moulin, and Bélanger (2010) propose a taxonomy of argumentation models that is horizontally divided into three categories: **micro-level** models, **macro-level** models, and **rhetorical** models.

In this article, we deal with argumentation on the **micro-level** (also called argumentation as a product or monological models). Micro-level argumentation focuses on the structure of a single argument. By contrast, **macro-level** models (also called *dialogical* models) and **rhetorical** models highlight the process of argumentation in a dialogue (Bentahar, Moulin, and Bélanger 2010, page 215). In other words, we examine the structure of a single argument produced by a single author in terms of its components, not the relations that can exist among arguments and their authors in time. A detailed discussion of these different perspectives can be found in Blair (2004), Johnson (2000), Reed and Walton (2003), Micheli (2011), O'Keefe (1982), and Rapanta, Garcia-Mila, and Gilabert (2013).[7]

2.2 Dimensions of Argument

The models mentioned here focus basically only on one dimension of the argument, namely, the logos dimension. According to the classical Aristotelian theory (Aristotle and Kennedy [translator] 1991), argument can exist in three dimensions, which are logos, pathos, and ethos. The **logos** dimension represents a proof by reason, an attempt to persuade by establishing a logical argument. For example, syllogism belongs to this argumentation dimension (Amossy 2009; Rapp and Wagner 2012). The **pathos** dimension makes use of appealing to emotions of the receiver and impacts its

7 There are, however, some argumentation theorists who disagree with this distinction and consider argumentation purely as dialogical. Freeman (2011) sees the argument as a process that is implicitly present even if the argumentation is a written text, which others treat as argument as product. For a deep discussion of opposing views on the dialectical nature of argumentation, we would point to Freeman (2011, page 53), Finocchiaro (2005), or to the pragma-dialectical approach by van Eemeren and Grootendorst (1984).

cognition (Micheli 2008). The **ethos** dimension of argument relies on the credibility of the arguer. This distinction will have practical impact later in Section 4.4, which deals with argumentation on the Web.

2.3 Original Toulmin Model

We conclude the theoretical section by presenting one (micro-level) argumentation model in detail: a widely used conceptual model of argumentation introduced by Toulmin (1958), which we will henceforth denote as **Toulmin original model**.[8] This model will play an important role later in the annotation studies (Section 4.4) and experimental work (Section 5.1). The model consists of six parts, referred to as **argument components**, where each component plays a distinct role.

Claim is an assertion put forward publicly for general acceptance (Toulmin, Rieke, and Janik 1984, page 29) or the conclusion we seek to establish by our arguments (Freeley and Steinberg 2008, page 153).

Data (Grounds) This is the evidence to establish the foundation of the claim (Schiappa and Nordin 2013) or, as simply put by Toulmin, "the data represent what we have to go on" (Toulmin 2003, page 90). The name of this concept was later changed to *grounds* in Toulmin, Rieke, and Janik (1984).

Warrant The role of *warrant* is to justify a logical inference from the *grounds* to the *claim*.

Backing is a set of information that stands behind the *warrant*. It assures its trustworthiness.

Qualifier limits the degree of certainty under which the argument should be accepted. It is the degree of force that the *grounds* confer on the *claim* in virtue of the *warrant* (Toulmin 2003, page 93).

Rebuttal presents a situation in which the *claim* might be defeated.

A schema of Toulmin's original model is shown in Figure 3. The lines and arrows symbolize implicit relations between the components. An example of an argument rendered using Toulmin's scheme can be seen in Figure 4.

We believe that this theoretical overview should provide sufficient background for the argumentation mining research covered in this article; for further references, we recommend, for example, van Eemeren et al. (2014).

3. Related Work in Computational Linguistics

We structure the related work into three sub-categories, namely, *argumentation mining*, *stance detection*, and *persuasion and on-line dialogs*, as these areas are closest to this article's focus. For a recent overview of general discourse analysis see (Webber, Egg, and Kordoni 2012). Apart from these, research on computer-supported argumentation has been also very active; see, for example, Scheuer et al. (2010) for a survey of various models and argumentation formalisms from the educational perspective or Schneider, Groza, and Passant (2013), who examine argumentation in the Semantic Web.

8 Henceforth, we will refer to the updated edition of Toulmin (1958), namely, Toulmin (2003).

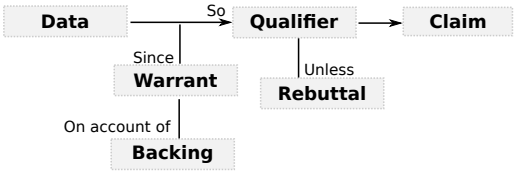

Figure 3
Original Toulmin's model of argument.

[Harry was born in Bermuda.]$_{\text{Data}}$ **Since** [A man born in Bermuda will generally be a British subject.]$_{\text{Warrant}}$ **On account of** [The following statuses and other legal provisions: (...)]$_{\text{Backing}}$ **So,** [presumably]$_{\text{Qualifier}}$ **Unless** [Both his parents were aliens]$_{\text{Rebuttal}}$ [Harry is a British subject.]$_{\text{Claim}}$

Figure 4
Example of an argument using Toulmin's model (Toulmin 2003).

3.1 Argumentation Mining

The argumentation mining field has been evolving very rapidly in recent years, resulting in several workshops co-located with major NLP conferences. We first present related work with a focus on annotations and then review experiments with classifying argument components, schemes, or relations.

3.1.1 Annotation Studies. One of the first papers dealing with annotating argumentative discourse was Argumentative Zoning for scientific publications (Teufel, Carletta, and Moens 1999). Later, Teufel, Siddharthan, and Batchelor (2009) extended the original seven categories to 15 and annotated 39 articles from two domains, where each sentence is assigned a category. The obtained Fleiss' κ were 0.71 and 0.65. In their approach, they tried to deliberately ignore the domain knowledge and rely only on general, rhetorical, and logical aspects of the annotated texts. In contrast to our work, argumentative zoning is specific to scientific publications and has been developed solely for that task.

Reed and Rowe (2004) presented *Araucaria*, a tool for argumentation diagramming that supports both convergent and linked arguments, missing premises (enthymemes), and refutations. They also released the *AracuariaDB* corpus, which has later been used for experiments in the argumentation mining field. However, the creation of the data set in terms of annotation guidelines and reliability is not reported—these limitations, as well as its rather small size, have been noted (Feng and Hirst 2011).

Biran and Rambow (2011) identified justifications for subjective claims in blog threads and Wikipedia talk pages. The data were annotated, with claims and their justifications reaching $\kappa = 0.69$, but a detailed description of the annotation approach was missing.

Schneider et al. (2013) annotated Wikipedia talk pages about deletion using 17 Walton's schemes (Walton 2007), reaching a moderate agreement (Cohen's $\kappa = 0.48$) and concluded that their analysis technique can be reused, although "it is intensive and difficult to apply."

Stab and Gurevych (2014a) annotated 90 argumentative essays (about 30k tokens), annotating claims, major claims, and premises and their relations (support, attack). They

reached Krippendorff's $\alpha_U = 0.72$ for argument components and Krippendorff's $\alpha = 0.81$ for relations between components.

Rosenthal and McKeown (2012) annotated sentences that are opinionated claims, in which the author expresses a belief that should be adopted by others. Two annotators labeled sentences as claims without any context and achieved Cohen's $\kappa = 0.50$ (2,000 sentences from LiveJournal) and 0.56 (2,000 sentences from Wikipedia).

Aharoni et al. (2014) performed an annotation study in order to find context-dependent claims and three types of context-dependent evidence in Wikipedia that were related to 33 controversial topics. The claim and evidence were annotated in 104 articles. The average Cohen's κ between a group of 20 expert annotators was 0.40. Compared with our work, the linguistic properties of Wikipedia are qualitatively different from other user-generated content, such as blogs or user comments (Ferschke 2014).

Wacholder et al. (2014) annotated "argument discourse units" in blog posts and criticized the Krippendorff's α_U measure. They proposed a new inter-annotator metric by taking the most overlapping part of one annotation as the "core" and all annotations as a "cluster." The data were extended by Ghosh et al. (2014), who annotated "targets" and "callouts" on the top of the units.

Park and Cardie (2014) annotated about 10k sentences from 1,047 documents into four types of argument propositions with Cohen's $\kappa = 0.73$ on 30% of the data set. Only 7% of the sentences were found to be non-argumentative.

Faulkner (2014) used Amazon Mechanical Turk to annotate 8,179 sentences from student essays. Three annotators decided whether the given sentence offered reasons for or against the main prompt of the essay (or no reason at all; 66% of the sentences were found to be neutral and easy to identify). The achieved Cohen's κ was 0.70.

The research has also been active on non-English data sets. Goudas et al. (2014) focused on user-generated Greek texts. They selected 204 documents and manually annotated sentences that contained an argument (760 out of 16,000). They distinguished claims and premises, but the claims were always implicit. However, the annotation agreement was not reported; neither was the number of annotators or the guidelines. A study on annotation of arguments was conducted by Peldszus and Stede (2013b), who evaluate agreement among 26 "naive" annotators (annotators with very little training). They manually constructed 23 German short texts, each of them containing exactly one central claim, two premises, and one objection (rebuttal or undercut), and analyzed annotator agreement on this artificial data set. Peldszus (2014) later achieved higher inter-rater agreement with expert annotators on an extended version of the same data. Kluge (2014b) built a corpus of argumentative German Web documents, containing 79 documents from seven educational topics, which were annotated by three annotators according to the claim-premise argumentation model. The corpus comprises 70,000 tokens and the inter-annotator agreement was 0.40 (Krippendorff's α_U). Houy et al. (2013) targeted argumentation mining of German legal cases.

Table 1 gives an overview of annotation studies with their respective argumentation model, domain, size, and agreement. It also contains other studies outside of computational linguistics and a few proposals and position papers.

3.1.2 Argument Analysis. Arguments in the legal domain were targeted in Mochales and Moens (2011). Using argumentation formalism inspired by Walton (2012), they used a multinomial naive Bayes classifier and maximum entropy model for classifying argumentative sentences on the *AraucariaDB* corpus (Reed and Rowe 2004). The same test data set was used by Feng and Hirst (2011), who utilized the C4.5 decision classifier. Rooney, Wang, and Browne (2012) investigated the use of convolution kernel methods

Table 1
Previous works on annotating argumentation. IAA = Inter-annotator agreement; N/A = not applicable.

Source	Arg. Model	Domain	Size	IAA
Newman and Marshall (1991)	Toulmin (1958)	legal domain (People vs. Carney, U.S. Supreme Court)	qualitative	N/A
Bal and Dizier (2010)	proprietary	socio-political newspaper editorials	56 documents	Cohen's κ (0.80)
Feng and Hirst (2011)	Walton, Reed, and Macagno (2008) (top 5 schemes)	legal domain (AracuariaDB corpus, 61% subset annotated with Walton scheme)	$\approx$ 400 arguments	not reported claimed to be small
Biran and Rambow (2011)	proprietary	Wikipedia Talk pages, blogs	309 + 118	Cohen's κ (0.69)
Georgila et al. (2011)	proprietary	general discussions (negotiations between florists)	21 dialogs	Krippendorff's α (0.37-0.56)
Mochales and Moens (2011)	Claim-Premise based on Freeman (1991)	legal domain (AracuariaDB corpus, European Human Rights Council)	641 documents w/ 641 arguments (AracuariaDB) 67 documents w/ 257 arguments (EHRC)	not reported
Walton (2012)	Walton, Reed, and Macagno (2008) (14 schemes)	political argumentation	256 arguments	not reported
Rosenthal and McKeown (2012)	opinionated claim, sentence level	blog posts, Wikipedia discussions	4000 sentences	Cohen's κ (0.50-0.57)
Conrad, Wiebe, and Hwa (2012)	proprietary (spans of arguing subjectivity)	editorials and blog post about ObamaCare	84 documents	Cohen's κ (0.68) on 10 documents
Schneider and Wyner (2012)	proprietary, argumentation schemes	camera reviews	N/A (proposal/position paper)	N/A
Schneider, Davis, and Wyner (2012)	Dung (1995) + Walton, Reed, and Macagno (2008)	unspecified social media	N/A (proposal/position paper)	N/A
Villalba and Saint-Dizier (2012)	proprietary, RST	hotel reviews, hi-fi products, political campaign	50 documents	not reported
Peldszus and Stede (2013a)	Freeman (1991) + RST	Potsdam Commentary Corpus	N/A (proposal/position paper)	N/A
Florou et al. (2013)	none	public policy making	69 argumentative segments / 322 non-argumentative segments	not reported
Peldszus and Stede (2013b)	based on Freeman (1991)	not reported, artificial documents created for the study	23 short documents	Fleiss' κ multiple results
Sergeant (2013)	N/A	Car Review Corpus (CRC)	N/A (proposal/position paper)	N/A
Wachsmuth et al. (2014)	none	hotel reviews	2100 reviews	Fleiss' κ (0.67)
Procter, Vis, and Voss (2013)	proprietary (Claim, Counter-claim)	Riot Twitter Corpus	7729 tweets under 'Rumors' category	percentage agreement (89% – 96%)
Stab and Gurevych (2014a)	Claim-Premise based on Freeman (1991)	student essays	90 documents	Kripp. α_U (0.72) Kripp. α (0.81)
Aharoni et al. (2014)	proprietary (claims, evidence)	Wikipedia	104 documents	Cohen's κ (0.40)
Park and Cardie (2014)	proprietary (argument propositions)	policy making (passenger rights and consumer protection)	1047 documents	Cohen's κ (0.73)
Goudas et al. (2014)	proprietary (premises)	social media	204 documents	not reported
Faulkner (2014)	none ("supporting argument")	student essays	8176 sentences	Cohen's κ (0.70)

for classifying whether a sentence belongs to an argumentative element or not, using the same corpus.

Stab and Gurevych (2014b) classified sentences to four categories (none, major claim, claim, premise) using their previously annotated corpus (Stab and Gurevych 2014a) and reached a 0.72 macro-F_1 score. In contrast to our work, their documents are expected to comply with a certain structure of argumentative essays and are assumed to always contain argumentation.

Biran and Rambow (2011) identified justifications on the sentence level using a naive Bayes classifier over a feature set based on statistics from the RST Treebank, namely, *n*-grams that were manually processed by deleting *n*-grams that "seemed irrelevant, ambiguous or domain-specific."

Llewellyn et al. (2014) experimented with classifying tweets into several argumentative categories, namely, claims and counter-claims (with and without evidence) and verification inquiries previously annotated by Procter, Vis, and Voss (2013). They used unigrams, punctuation, and POS as features in three classifiers.

Park and Cardie (2014) classified propositions into three classes (unverifiable, verifiable non-experimental, and verifiable experimental) and ignored non-argumentative texts. Using multi-class support vector machine (SVM) and a wide range of features (*n*-grams, POS, sentiment clue words, tense, person) they achieved Macro$F_1 = 0.69$.

Peldszus (2014) experimented with a rather complex labeling schema of argument segments, but the data were artificially created for the task and manually cleaned (such as removing segments that did not meet the criteria or non-argumentative segments).

In the first step of their two-phase approach, Goudas et al. (2014) sampled the data set to be balanced and identified argumentative sentences with $F_1 = 0.77$, using the maximum entropy classifier. For identifying premises, they used BIO encoding of tokens and achieved an F_1 score of 0.42 using conditional random fields (CRFs).

Saint-Dizier (2012) developed a Prolog engine using a lexicon of 1,300 words and a set of 78 hand-crafted rules with the focus on a particular argument structure, "reasons supporting conclusions," in French.

Taking the dialogical perspective, Cabrio and Villata (2012) built upon an argumentation framework proposed by Dung (1995), which models arguments within a graph structure and provides a reasoning mechanism for resolving accepted arguments. For identifying support and attack, they relied on existing research on textual entailment (Dagan et al. 2009), namely, using the off-the-shelf *EDITS* system. The test data were taken from a debate portal *Debatepedia* and covered 19 topics. Evaluation was performed in terms of measuring the acceptance of the "main argument" using the automatically recognized entailments, yielding an F_1 score of about 0.75. In contrast to our work, which deals with micro-level argumentation, Dung's model is an abstract framework intended to model dialogical argumentation.

Finding a bridge between existing discourse research and argumentation has been targeted by several researchers. Peldszus and Stede (2013a) surveyed literature on argumentation and proposed utilization of Rhetorical Structure Theory (RST) (Mann and Thompson 1987). They claimed that RST is by its design well-suited for studying argumentative texts, but an empirical evidence has not yet been provided. Penn Discourse Tree Bank (PDTB) (Prasad et al. 2008) relations have been under examination by argumentation mining researchers too. Cabrio, Tonelli, and Villata (2013) examined a connection between five of Walton's schemes and discourse markers in PDTB: however, an empirical evaluation is missing.

3.2 Stance Detection

Research related to argumentation mining also involves **stance detection**. In this case, the whole document (discussion post, article) is assumed to represent the writer's standpoint on the discussed topic. Because the topic is stated as a controversial question, the author is either *for* or *against* it.

Somasundaran and Wiebe (2009) built a computational model for recognizing stances in dual-topic debates about named entities in the electronic products domain by combining preferences learned from the Web data and discourse markers from PDTB (Prasad et al. 2008). Hasan and Ng (2013) determined stance in online ideological debates on four topics using data from createdebate.com, utilizing supervised machine learning and features ranging from *n*-grams to semantic frames. Predicting stance of posts in *Debatepedia* as well as external articles using a probabilistic graphical model was presented in Gottipati et al. (2013). This approach also used sentiment lexicons and Named Entity Recognition as a preprocessing step and achieved accuracy about 0.80 in binary prediction of stances in debate posts.

Recent research has involved joint modeling, taking into account information about the users, the dialog sequences, and others. Hasan and Ng (2012) proposed a machine learning approach to debate stance classification by leveraging contextual information and author's stances towards the topic. Qiu, Yang, and Jiang (2013) introduced a computational debate side model to cluster posts or users by sides for general threaded discussions using a generative graphical model utilizing words from various subjectivity lexicons as well as all adjectives and adverbs in the posts. Qiu and Jiang (2013) proposed a graphical model for viewpoint discovery in discussion threads. Burfoot, Bird, and Baldwin (2011) exploited the informal citation structure in U.S. Congressional floor-debate transcripts and used a collective classification, which outperforms methods that consider documents in isolation.

Some works also utilize argumentation-motivated features. Park, Lee, and Song (2011) dealt with contentious issues in Korean newswire discourse. Although they annotate the documents with "argument frames," the formalism remains unexplained and does not refer to any existing research in argumentation. Walker et al. (2012) incorporated features with some limited aspects of the argument structure, such as cue words signaling rhetorical relations between posts, POS generalized dependencies, and a representation of the parent post (context) to improve stance classification over 14 topics from convinceme.net.

3.3 Online Persuasion

Another stream of research has been devoted to persuasion in online media, which we consider as a more general research topic than argumentation.

Schlosser (2011) investigated persuasiveness of online reviews and concluded that presenting two sides is not always more helpful and can even be less persuasive than presenting one side. Mohammadi et al. (2013) explored persuasiveness of speakers in YouTube videos and concluded that people are perceived more persuasive in video than in audio and text. Miceli, de Rosis, and Poggi (2006) proposed a computational model that attempts to integrate emotional and non-emotional persuasion. In the study of Murphy (2001), persuasiveness was assigned to 21 articles (out of 100 manually preselected) and four of them were later analyzed in detail for comparing the perception of persuasion between experts and students. Bernard, Mercier, and Clément (2012) experimented with children's perception of discourse connectives (namely, with *because*)

to link statements in arguments and found out that 4- and 5-year-olds and adults are sensitive to the connectives. Le (2004) presented a study of persuasive texts and argumentation in newspaper editorials in French.

A coarse-grained view on dialogs in social media was examined by Bracewell, Tomlinson, and Wang (2013), who proposed a set of 15 social acts (such as agreement, disagreement, or supportive behavior) to infer the social goals of dialog participants and presented a semi-supervised model for their classification. Their social act types were inspired by research in psychology and organizational behavior and were motivated by work in dialog understanding. They annotated a corpus in three languages using in-house annotators and achieved κ in the range from 0.13 to 0.53.

Georgila et al. (2011) focused on cross-cultural aspects of persuasion or argumentation dialogs. They developed a novel annotation scheme stemming from different literature sources on negotiation and argumentation as well as from their original analysis of the phenomena. The annotation scheme is claimed to cover three dimensions of an utterance, namely, speech act, topic, and response or reference to a previous utterance. They annotated 21 dialogs and reached Krippendorff's α between 0.38 and 0.57.

Summary of Related Work Section. Given the broad landscape of various approaches to argument analysis and persuasion studies presented in this section, we would like to stress some novel aspects of the current article. First, we aim to adapt a model of argument based on research by argumentation scholars, both theoretical and empirical. We pose several pragmatical constraints, such as register independence (generalization over several registers). Second, our emphasis is on reliable annotations and sufficient data size (about 90k tokens). Third, we deal with fairly unrestricted Web-based sources, so additional steps of distinguishing whether the texts are argumentative are required. Argumentation mining has been a rapidly evolving field with several major venues in 2015. We encourage readers to consult an upcoming survey article by Lippi and Torroni (2016) or the proceedings of the Second Argumentation Mining workshop (Cardie 2015) to keep up with recent developments. However, to the best of our knowledge, the main findings of this article have not yet been made obsolete by any related work.

4. Annotation Studies and Corpus Creation

This section describes the process of data selection, annotation, curation, and evaluation with the goal of creating a new corpus suitable for argumentation mining research in the area of computational linguistics. As argumentation mining is an evolving discipline without established and widely accepted annotation schemes, procedures, and evaluation, we want to keep this overview detailed to ensure full reproducibility of our approach. Given the wide range of perspectives on argumentation itself (van Eemeren et al. 2014), variety of argumentation models (Bentahar, Moulin, and Bélanger 2010), and high costs of discourse or pragmatic annotations (Prasad et al. 2008), creating a new, reliable corpus for argumentation mining represents a substantial effort.

A motivation for creating a new corpus stems from the various use-cases discussed in the introduction, as well as some research gaps pointed out in Section 1 and further discussed in the survey in Section 3.1 (e.g., domain restrictions, missing connection to argumentation theories, non-reported reliability or detailed schemes).

4.1 Topics and Registers

As a main field of interest in the current study, we chose controversies in education. One distinguishing feature of educational topics is their breadth of sub-topics and points of view, as they attract researchers, practitioners, parents, students, and policy-makers. We assume that this diversity leads to the linguistic variability of the education topics and thus represents a challenge for NLP. In cooperation with researchers from the German Institute for International Educational Research,[9] we identified the following current controversial topics in education in English-speaking countries: (1) **homeschooling**; (2) **public versus private schools**; (3) **redshirting**—intentionally delaying the entry of an age-eligible child into kindergarten, allowing their child more time to mature emotionally and physically (Huang and Invernizzi 2013); (4) **prayer in schools**—whether prayer in schools should be allowed and taken as a part of education or banned completely; (5) **single-sex education**—single-sex classes (boys and girls separate) versus mixed-sex classes ("co-ed"); and (6) **mainstreaming**—including children with special needs into regular classes.

Because we were also interested in whether argumentation differs across registers,[10] we included four different registers, namely, (1) user **comments** to newswire articles or to blog posts; (2) posts in discussion forums (**forum posts**); (3) **blog posts**; and (4) newswire **articles**.[11] Throughout this work, we will refer to each article, blog post, comment, or forum posts as a **document**. This variety of sources covers mainly user-generated content except newswire articles, which are written by professionals and undergo an editing procedure by the publisher. Because many publishers also host blog-like sections on their portals, we consider as blog posts all content that is hosted on personal blogs or clearly belong to a blog category within a newswire portal.

4.2 Raw Corpus Statistics

Given the six controversial topics and four different registers, we compiled a collection of plain-text documents, which we call the **raw corpus**. It contains 694,110 tokens in 5,444 documents. As a coarse-grained analysis of the data, we examined the lengths and the number of paragraphs (see Figure 5). Comments and forum posts follow a similar distribution, being shorter than 300 tokens on average. By contrast, articles and blogs are longer than 400 tokens and have 9.2 paragraphs on average. The process of compiling the raw corpus and its further statistics are described in detail in Appendix A.

4.3 Annotation Study 1: Identifying Persuasive Documents in Forums and Comments

The goal of this study was to select documents suitable for a fine-grained analysis of arguments. In a preliminary study on annotating argumentation using a small

9 `http://www.dipf.de`.

10 The distinction between registers is based on the situational context and the functional characteristics (Biber and Conrad 2009, page 6).

11 We ignored social media sites and micro-blogs, either because searching and harvesting data is technically challenging (Facebook, Google Plus), or the texts are too short to convey argumentation, as seen in our preliminary experiments (the case of Twitter). We also did not consider debate portals (sites with *pros* and *cons* threads). We observed that they contain many artificial controversies or non-sense topics (for instance, createdebate.com) or their content is professionally curated (idebate.org, for example). However, we admit that debate portals might be a valuable resource in the argumentation mining research.

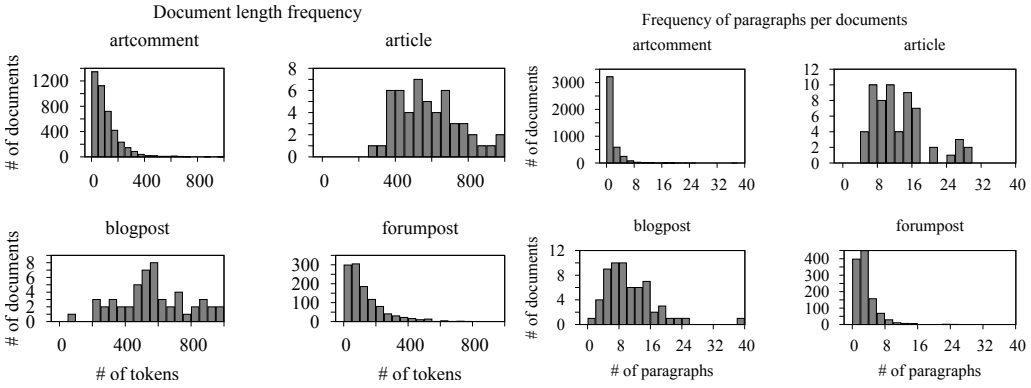

Figure 5
Number of documents with a certain number of tokens (left) and paragraphs (right) in the *raw corpus*.

sample (50 random documents) of forum posts and comments from the raw corpus, we found that many documents convey no argumentation at all, even in discussions about controversies. We observed that such contributions do not intend to persuade; these documents typically contain story-sharing, personal worries, user interaction (asking questions, expressing agreement), off-topic comments, and others. Such characteristics are typical of online discussions in general, but they have not been examined with respect to argumentation or persuasion. Indeed, we observed that there are (1) documents that are completely unrelated and (2) documents that are related to the topic, but do not contain any argumentation. This issue has been identified among argumentation theorists—for example, as *external relevance* by Paglieri and Castelfranchi (2014). Similar findings were also confirmed in related literature in argumentation mining, although never tackled empirically (Garcia-Villalba and Saint-Dizier 2014; Park and Cardie 2014). These documents are thus not suitable for analyzing argumentation.

In order to filter documents that are suitable for argumentation annotation, we defined a binary document-level classification task. The distinction is made between either **persuasive documents** or **non-persuasive** (which includes all other sorts of texts, such as off-topic, story sharing, unrelated dialog acts, etc.).

4.3.1 Annotation Study. The two annotated categories were *on-topic persuasive* and *non-persuasive*.[12] Three annotators with near-native English proficiency annotated a set of 990 documents (a random subset of comments and forum posts) reaching a Fleiss' κ of 0.59. The final label was selected by majority voting. The annotation study took an average of 15 hours per annotator, with approximately 55 annotated documents per hour. The resulting labels were derived by majority voting. Out of 990 documents, 524 (53%) were labeled as on-topic persuasive. We will refer to this corpus as **gold data persuasive**.

Sources of Disagreement. We examined all disagreements between annotators and discovered some typical problems, such as implicitness or topic relevance. First, the

12 We also initially experimented with three to five categories using a Likert scale but found no extra benefits over the binary decision and thus decided to keep only two categories after the pilot experiments.

authors often express their stance towards the topic *implicitly*, so it must be inferred by the reader. To do so, certain common-ground knowledge is required. However, such knowledge heavily depends on many aspects, such as the reader's familiarity with the topic or her cultural background, as well as the context of the source Web site or the discussion forum thread. This also applies for sarcasm and irony. Second, the decision whether a particular topic is *persuasive* was always made with respect to the controversial topic under examination. Some authors *shift the focus* to a particular aspect of the given controversy or a related issue, making the document less relevant.

4.3.2 Discussion. We achieved moderate[13] agreement between the annotators, although the definition of persuasiveness annotation might seem a bit fuzzy.[14] We found different amounts of persuasion in the specific topics. For instance, *prayer in schools* or *private vs. public schools* attract persuasive discourse, while other discussed controversies often contain non-persuasive discussions (represented by *redshirting* and *mainstreaming*). Although these two topics are also highly controversial, the participants of online discussions seem to not attempt to persuade but they rather exchange information, support others in their decisions, and so forth. This was also confirmed by socio-psychological researchers. Ammari, Morris, and Schoenebeck (2014) show that parents of children with special needs rely on discussion sites for accessing information and social support and that, in particular, posts containing humor, achievement, or treatment suggestions are perceived to be more socially appropriate than posts containing judgment, violence, or social comparisons. According to Nicholson and Leask (2012), in the online forum, parents of autistic children were seen to understand the issue because they had lived it. Assuming that participants in discussions related to young children (e.g., redshirting or mainstreaming) are usually women (mothers), gender can also play a role. In a study of online persuasion, Guadagno and Cialdini (2002) conclude that women chose to bond rather than compete (women feel more comfortable cooperating, even in a competitive environment), whereas men are motivated to compete if necessary to achieve independence.

4.4 Annotation Study 2: Annotating Micro-structure of Arguments

The goal of this study was to annotate documents on a detailed level with respect to an argumentation model. First, we will present the annotation scheme. Second, we will

13 Following the terminology proposed by Landis and Koch (1977, page 165), although they claim that the *divisions are clearly arbitrary.* For a detailed discussion on interpretation of agreement values, see for example Artstein and Poesio (2008).

14 We also experimented with different task definition before in the preliminary studies. However, *identifying argumentative documents* was misleading, as the annotators expected a reasonable argument. For instance, consider the following example: *Doc#1247 (artcomment, prayer-in-schools): Keep church and state separate. Period.* This is not an argumentative text in the traditional sense of giving reason, however, the persuasion is obvious. We are interested in all kinds of persuasive documents, not only in those that contain some clearly defined argument structures, as they can still contain useful information for decision-making. Trabelsi and Zaiane (2014) defined a *contentious document* as a document that contains expressions of one or more divergent viewpoints in response to the contention question but they did not tackle the classification of these documents. Our task also resembles aspect-based sentiment analysis (ABSA), where the aspect in our case would be the controversial topic. However, in contrast to the research in ABSA, the aspects in our case are purely abstract entities and current approaches to model ABSA do not clearly fit our task.

describe the annotation process. Finally, we will evaluate the agreement and draw some conclusions.

4.4.1 Argumentation Model Selection. Given the theoretical background briefly introduced in Section 2, we motivate our selection of the argumentation model with the following requirements. First, the scope of this work is to capture argumentation within a single document, thus focusing on micro-level models. Second, there should exist empirical evidence that such a model has been used for analyzing argumentation in previous works, so it is likely to be suitable for our purpose of argumentative discourse analysis in user-generated content. Regarding the first requirement, two typical examples of micro-level models are *Toulmin's model* (Toulmin 1958) and *Walton's schemes* (Walton, Reed, and Macagno 2008). Let us now elaborate on the second requirement.

Walton's Schemes. Walton's argumentation schemes are claimed to be general and domain independent. Nevertheless, evidence from the computational linguistics field shows that the schemes lack coverage for analyzing real argumentation in natural language texts. In examining real-world political argumentation from Walton (2005), Walton (2012) found that 37.1% of the arguments collected did not fit any of the 14 schemes he had chosen so he created new schemes ad hoc. Cabrio, Tonelli, and Villata (2013) selected five argumentation schemes from Walton and mapped these patterns to discourse relation categories in the Penn Discourse TreeBank (PDTB) (Prasad et al. 2008), but later they had to define two new argumentation schemes that they discovered in PDTB. Similarly, Ong, Litman, and Brusilovsky (2014) admitted that the schemes are ambiguous and hard to directly apply for annotation, therefore they modified the schemes and created new ones that matched the data.

Although Macagno and Konstantinidou (2012) show several examples of two argumentation schemes applied to a few selected arguments in classroom experiments, empirical evidence presented by Anthony and Kim (2014) reveals many practical and theoretical difficulties of annotating dialogues with schemes in classroom deliberation, providing many details on the arbitrary selection of the subset of the schemes and the ambiguity of the scheme definitions, concluding that the presence of the authors during the experiment was essential for inferring and identifying the argument schemes (Anthony and Kim 2014, page 93).

Toulmin Model. Although this model (refer to Section 2.3) was designed to be applicable to real-life argumentation, there are numerous studies criticizing both the clarity of the model definition and the differentiation between elements of the model. Ball (1994) claims that the model can be used only for the most simple arguments and fails on the complex ones. Additionally, Freeman (1991) and other argumentation theorists criticize the usefulness of Toulmin's framework for the description of real-life argumentative texts. However, others have advocated the model and claimed that it can be applied to the people's ordinary argumentation (Simosi 2003; Dunn 2011).

A number of studies (outside the field of computational linguistics) used Toulmin's model as their backbone argumentation framework. Chambliss (1995) experimented with analyzing 20 written documents in a classroom setting in order to find the argument patterns and parts. Simosi (2003) examined employees' argumentation to resolve conflicts. Voss (2006) analyzed experts' protocols dealing with problem-solving.

The model has also been used in research on computer-supported collaborative learning. Erduran, Simon, and Osborne (2004) adapt Toulmin's model for coding classroom argumentative discourse among teachers and students. Stegmann et al. (2011)

build on a simplified Toulmin model for scripted construction of arguments in computer-supported collaborative learning. Garcia-Mila et al. (2013) coded utterances into categories from Toulmin's model in persuasion and consensus-reaching among students. Weinberger and Fischer (2006) analyze asynchronous discussion boards in which learners engage in an argumentative discourse with the goal of acquiring knowledge. For coding the argument dimension, they created a set of argumentative moves based on Toulmin's model. Given this empirical evidence, we decided to build upon Toulmin's model.

4.4.2 Adaptation of Toulmin's Model of Argumentation in User-generated Web Discourse. In this annotation task, a sequence of tokens (e.g., a phrase, a sentence, or any arbitrary text span) is labeled with a corresponding argument component (such as the *claim*, the *grounds*, and others). There are no explicit relations between these annotation spans as the relations are implicitly encoded in the pragmatic function of the components in Toulmin's model.

In order to prove the suitability of Toulmin's model, we analyzed 40 random documents from the *gold data persuasive* data set using the original Toulmin model as presented in Section 2.3.[15] We took into account several criteria for assessment, such as frequency of occurrence of the components or their importance for the task. We proposed some modifications of the model based on the following observations.

No Qualifier. Authors do not state the degree of cogency (the probability of their *claim*, as proposed by Toulmin). Thus we omitted *qualifier* from the model because of its absence in the data.

No Warrant. The **warrant** as a logical explanation why one should accept the claim given the evidence is almost never stated. As pointed out by Toulmin (2003, page 92), "data are appealed to explicitly, warrants implicitly." This observation has also been made by Voss (2006). Also, according to van Eemeren, Grootendorst, and Kruiger (1987, page 205), the distinction of warrant is perfectly clear only in Toulmin's examples, but the definitions fail in practice. We omitted *warrant* from the model.

Attacking the Rebuttal. **Rebuttal** is a statement that attacks the *claim*, thus playing a role of an opposing view. In reality, the authors often attack the presented rebuttals by another counter-rebuttal in order to keep the whole argument's position consistent. Thus we introduced a new component—**refutation**—which is used for attacking the *rebuttal*. Annotation of *refutation* was conditioned by the explicit presence of *rebuttal* and enforced by the annotation guidelines. The chain *rebuttal–refutation* is also known as the **procatalepsis** figure in rhetoric, in which the speaker raises an objection to his own argument and then immediately answers it. By doing so, the speaker hopes to strengthen the argument by dealing with possible counter-arguments before the audience can raise them (Walton 2007, page 106).

Implicit Claim. The claim of the argument should always reflect the main standpoint with respect to the discussed controversy. We observed that this standpoint is not always

15 The reason we are focusing on comments and forum posts in the first place is pragmatical; Kluge (2014a) abstained from Toulmin's model when annotating long German newswire documents because of high time costs. Nevertheless, we will include another register later in our experiments (Section 4.4.4).

explicitly expressed, but remains implicit and must be inferred by the reader. Therefore, we allow the claim to be implicit. In such a case, the annotators must explicitly write down the (inferred) stance of the author.

Multiple Arguments in One Document. By definition, Toulmin's model is intended to model a single argument, with the *claim* in its center. However, we observed in our data that some authors elaborate on both sides of the controversy equally and put forward an argument for each side (by *argument* here we mean the *claim* and its *premises*, *backings*, etc.). Therefore we allow multiple arguments to be annotated in one document. At the same time, we restrained the annotators from creating complex argument hierarchies.

Terminology. Toulmin's *grounds* have an equivalent role to a *premise* in the classical view of an argument (Reed and Rowe 2006; van Eemeren et al. 2014) in terms that they offer the reasons why one should accept the standpoint expressed by the *claim*. As this terminology has been used in several related works in the argumentation mining field (Mochales and Moens 2011; Peldszus and Stede 2013b; Ghosh et al. 2014; Stab and Gurevych 2014a), we will keep this convention and denote the *grounds* as *premises*.

The Role of Backing. One of the main critiques of the original Toulmin model was the vague distinction between *grounds*, *warrant*, and *backing* (Freeman 1991; Newman and Marshall 1991; Hitchcock 2003). The role of *backing* is to give additional support to the *warrant*, but there is no *warrant* in our model anymore. However, what we observed during the analysis was the presence of some *additional evidence*. Such evidence does not play the role of the *grounds* (*premises*) as it is not meant as a reason supporting the *claim*, but it also does not explain the reasoning, thus is not a *warrant* either. It usually supports the whole argument and is stated by the author as a certain fact. Therefore, we extended the scope of *backing* as an additional support to the whole argument.

The annotators were instructed to distinguish between *premises* and *backing*, so that *premises* should cover generally applicable reasons for the claim, whereas *backing* is a single personal experience or statements that give credibility or attribute certain expertise to the author. As a sanity check, the argument should still make sense after removing *backing* (would only be considered "weaker").

4.4.3 Model Definition. We call the model a **modified Toulmin model**. It contains five argument components, namely, *claim*, *premise*, *backing*, *rebuttal*, and *refutation*. When annotating a document, any arbitrary token span can be labeled with an argument component; the components do not overlap. The spans are not known in advance and the annotator thus chooses the span and the component type at the same time. All components are optional (they do not have to be present in the argument) except the *claim*, which is either explicit or implicit (see above). If a token span is not labeled by any argument component, it is not considered as a part of the argument and is later denoted as *none* (this category is not assigned by the annotators).

An example analysis of a forum post is shown in Figure 6. Figure 7 then shows a diagram of the analysis from that example (the content of the argument components was shortened or rephrased).

4.4.4 Annotation Workflow. The annotation experiment was split into three phases. All documents were annotated by three independent annotators, who participated in two training sessions. During the first phase, 50 random comments and forum posts were annotated. Problematic cases were resolved after discussion and the guidelines were

Doc#4733 (forumpost, public-private-schools) [*claim:* The public schooling system is not as bad as some may think.] [*rebuttal:* Some mentioned that those who are educated in the public schools are less educated,] [*refutation:* well I actually think it would be in the reverse.] [*premise:* Student who study in the private sector actually pay a fair amount of fees to do so and I believe that the students actually get let off for a lot more than anyone would in a public school. And its all because of the money.¶
In a private school, a student being expelled or suspended is not just one student out the door, its the rest of that students schooling life fees gone. Whereas in a public school, its just the student gone.]¶
[*backing:* I have always gone to public schools and when I finished I got into University. I do not feel disadvantaged at all.]

Figure 6
An annotation example using the *modified Toulmin model*.

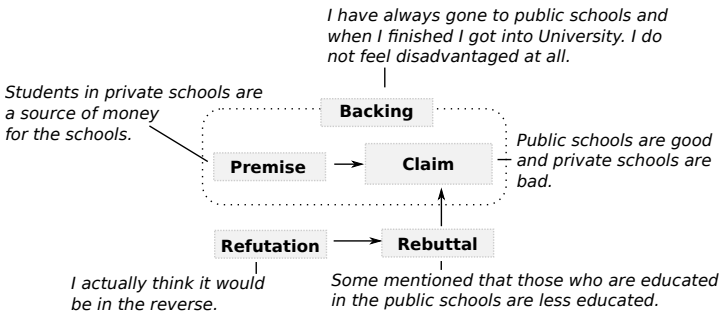

Figure 7
Modified Toulmin's model used for annotation of arguments with an instantiated example from a single discussion forum post on *public vs. private schools* (see Figure 6). The arrows show relations between argument components; the relations are implicit and inherent in the model. By contrast to the example of the *original Toulmin model* in Figure 3, we do not propose any connective phrases to the relations (such as *so*, *unless*, etc.).

refined. In the second phase, we wanted to extend the range of annotated registers, so we selected 148 comments and forum posts as well as 41 blog posts. After the second phase, the annotation guidelines were final.[16]

In the final phase, we extended the range of annotated registers and added newswire articles from the *raw corpus* in order to test whether the annotation guidelines (and inherently the model) was general enough. Therefore we selected 96 comments/forum posts, 8 blog posts, and 8 articles for this phase. A detailed inter-annotator agreement study on documents from this final phase is reported in Section 4.4.6.

The annotations were very time-consuming. In total, each annotator spent 35 hours annotating over the course of five weeks. Discussions and consolidation of the gold data took another 6 hours. Comments and forum posts required an average of 4 minutes per document to annotate, whereas blog posts and articles required an average of 14 minutes per document. Examples of annotated documents from the gold data are listed in Appendix A.

16 The annotation guidelines are available under CC-BY-SA license at `https://www.ukp.tu-darmstadt.de/data/argumentation-mining/`.

Table 2
Topic and register distribution in the *gold standard Toulmin* corpus.

Topic \ Register	Comment	Forum post	Blog post	Article	Total
Homeschooling	32	12	11	1	56
Mainstreaming	12	5	3	1	21
Prayer in schools	31	14	10	0	55
Public vs. private	117	10	7	0	134
Redshirting	19	13	4	1	37
Single-sex education	14	12	9	2	37
Total	216	73	46	5	340

Table 3
Gold standard Toulmin corpus statistics.

Register	Tokens	Mean	Sentences	Mean
Comments	35,461	164.17 ± 155.87	1,748	8.09 ± 7.68
Forum posts	13,033	178.53 ± 132.33	641	8.78 ± 7.53
Blogs	32,731	711.54 ± 293.72	1,378	29.96 ± 14.82
Articles	3,448	689.60 ± 183.34	132	24.60 ± 6.58
All	84,673	249.04 ± 261.77	3,899	11.44 ± 11.70

Discarding Documents During Annotation. We discarded 11 documents out of the total 351 annotated documents. Five forum posts, although annotated as *persuasive* in the first annotation study, were at a deeper look a mixture of two or more posts with missing quotations,[17] therefore unsuitable for analyzing argumentation. Three blog posts and two articles were found not to be argumentative (the authors took no stance to the discussed controversy) and one article was an interview, which the current model cannot capture (a dialogical argumentation model would be required).

For each of the 340 documents, the gold standard annotations were obtained using majority vote. If simple majority voting was not possible (different boundaries of the argument component together with a different component label), the gold standard was set after discussion among the annotators. We will refer to this corpus as the **gold standard Toulmin** corpus. The distribution of topics and registers in this corpus is shown in Table 2, and Table 3 presents some lexical statistics.

4.4.5 Annotation Set-up. Based on pre-studies, we set the minimal unit for annotation as *token*.[18] The documents were pre-segmented using the Stanford Core NLP sentence

17 All of them came from the same source Web site, which does not support any HTML formatting of quotations.

18 We also considered sentences or clauses. The sentence level seems to be reasonable in most of the cases, however, it is too coarse-grained if a sentence contains multiple clauses that belong to different argumentation components. Segmentation to clauses is not trivial and has been considered as a separate task since CoNLL 2001 (Tjong, Sang, and Déjean 2001). Best systems based on Join-CRF reach 0.81 F_1

splitter (Manning et al. 2014) embedded in the DKPro Core framework (Eckart de Castilho and Gurevych 2014). Annotators were asked to stick to the sentence level by default and label entire pre-segmented sentences. They should switch to annotations on the token level only if (a) a particular sentence contained more than one argument component, or (b) if the automatic sentence segmentation was wrong. Given the "noise" in user-generated Web data (wrong or missing punctuation, casing, etc.), this was often the case.

Annotators were also asked to rephrase (summarize) each annotated argument component into a simple statement when applicable, as shown in Figure 7. This was used as a first sanity checking step, as each argument component is expected to be a coherent discourse unit. For example, if a particular occurrence of a *premise* cannot be summarized/rephrased into one statement, this may require further splitting into two or more *premises*.

For the actual annotations, we developed a custom-made Web-based application that allowed users to switch between different granularity of argument components (tokens or sentences), to annotate the same document in different argument "dimensions" (logos and pathos), and to write a summary for each annotated argument component.

4.4.6 Inter-annotator Agreement. As a measure of annotation reliability, we rely on Krippendorff's unitized alpha (α_U) (Krippendorff 2004). To the best of our knowledge, this is the only agreement measure that is applicable when both *labels* and *boundaries* of segments are to be annotated.

Although the measure has been used in related annotation work (Ghosh et al. 2014; Kluge 2014a; Stab and Gurevych 2014a), there is one important detail that has not been properly communicated. The α_U is computed over a *continuum* of the smallest units, such as tokens. This continuum corresponds to a single document in the original Krippendorff's work. However, there are two possible extensions to multiple documents (a corpus), namely (a) to compute α_U for each document first and then report an average value, or (b) to concatenate all documents into one large continuum and compute α_U over it. The first approach with averaging yielded an extremely high standard deviation of α_U (i.e., avg. = 0.253; std. dev. = 0.886; median = 0.476 for the *claim*). This says that some documents are easy to annotate whereas others are harder, but interpretation of such an averaged value has no evidence either in Krippendorff (2004) or other papers based upon it. Thus we use the other methodology and treat the whole corpus as a single long continuum (which yields in the example of *claim* = α_U 0.541).[19]

Table 4 shows the inter-annotator agreement as measured on documents from the last annotation phase (see Section 4.4.4). The overall α_U for all register types, topics, and argument components is 0.48 in the *logos* dimension (annotated with the

score (Nguyen, Nguyen, and Shimazu 2009) for embedded clauses and 0.92 for non-embedded (Zhang et al. 2013). To the best of our knowledge, there is no available out-of-box solution for clause segmentation, thus we took *sentences* as another level of segmentation. Nevertheless, pre-segmenting the text to clauses and their relation to argument components deserves future investigation.

19 Another pitfall of the α_U measure when documents are concatenated to create a single continuum is that its value depends on the order of the documents (the annotated spans, respectively). We did the following experiment: Using 10 random annotated documents, we created all 362,880 possible concatenations and measured the α_U for each permutation. The resulting standard error was 0.002, so the influence of the ordering is rather low. Still, each reported α_U was averaged from 100 random concatenations of the analyzed documents.

Table 4
Inter-annotator agreement (Krippendorff's α_U) across various registers, topics, and argument components. **Bold** values emphasize $\alpha_U \geq 0.50$. Joint logos is a joint α_U for all argument components in the logos dimension (*claim*, *premise*, *backing*, *rebuttal*, *refutation*). HS = homeschooling; RS = redshirting; PIS = prayer in schools; SSE = single sex education; MS = mainstreaming; PPS = private vs. public schools.

	All topics	**HS**	**RS**	**PIS**	**SSE**	**MS**	**PPS**
(a) Comments + Forum posts							
Claim	**0.59**	**0.52**	0.36	**0.70**	**0.69**	**0.51**	**0.55**
Premise	**0.69**	0.35	0.31	**0.80**	0.47	0.16	0.38
Backing	0.48	0.15	0.06	0.36	**0.54**	0.14	0.49
Rebuttal	0.37	0.12	0.03	0.25	-0.02	**0.80**	0.34
Refutation	0.08	0.03	-0.01	-0.02	–	0.32	0.11
Joint logos	**0.60**	0.28	0.19	**0.68**	0.49	0.16	0.44
(b) Articles + Blog posts							
Claim	0.22	-0.02	-0.03	–	–	0.33	–
Premise	0.24	0.02	0.24	–	-0.04	0.40	–
Backing	-0.03	0.18	-0.20	–	0.26	-0.20	–
Rebuttal	0.01	–	0.08	–	-0.08	-0.08	–
Refutation	0.34	–	0.40	–	-0.01	-0.01	–
Joint logos	0.09	0.05	0.01	–	0.08	0.04	–
(c) Articles + Blog posts + Comments + Forum posts							
Claim	**0.54**	**0.52**	0.28	**0.70**	**0.71**	0.43	**0.55**
Premise	**0.62**	0.33	0.29	**0.80**	0.32	0.27	0.38
Backing	0.31	0.17	0.00	0.36	0.42	-0.04	0.49
Rebuttal	0.08	0.12	0.09	0.25	0.02	0.03	0.34
Refutation	0.17	0.03	0.38	-0.02	0.00	0.20	0.11
Joint logos	0.48	0.27	0.14	**0.68**	0.34	0.08	0.44

modified Toulmin model). Such agreement can be considered as moderate by the measures proposed by Landis and Koch (1977), however, direct interpretation of the agreement value lacks consensus (Artstein and Poesio 2008, page 591). Similar inter-annotator agreement numbers were achieved in the relevant work in argumentation mining (refer to Table 1 in Section 3.1; although most of the numbers are not directly comparable, as different inter-annotator metrics were used on different tasks).

There is a huge difference in α_U regarding the registers between comments + forums posts (α_U 0.60, Table 4a) and articles + blog posts (α_U 0.09, Table 4b) in the logos dimension. If we break down the value with respect to the individual argument components, the agreement on *claim* and *premise* is substantial in the case of comments and forum posts (0.59 and 0.69, respectively). By contrast, these argument components were annotated only with a fair agreement in articles and blog posts (0.22 and 0.24, respectively).

As can also be observed from Table 4, the annotation agreement in the logos dimension varies regarding the document topic. Whereas it is substantial/moderate for *prayer*

in schools (0.68) or *private vs. public schools* (0.44), for some topics it remains rather slight, such as in the case of *redshirting* (0.14) or *mainstreaming* (0.08).

4.4.7 Causes of Disagreement—Quantitative Analysis. First, we examine the disagreement in annotations by posing the following research question: *Are there any measurable properties of the annotated documents that might systematically cause low inter-annotator agreement?* We use Pearson's correlation coefficient between α_U on each document and the particular property under investigation. We investigated the following set of measures.

- *Full sentence coverage ratio* represents a ratio of argument component boundaries that are aligned to sentence boundaries. The value is 1.0 if all annotations in the particular document are aligned to sentences and 0.0 if no annotations match the sentence boundaries. Our hypothesis was that automatic segmentation to sentences was often incorrect, therefore annotators had to switch to the token-level annotations and this might have increased disagreement on boundaries of the argument components.
- *Document length*, *paragraph length*, and *average sentence length*. Our hypotheses was that the length of documents, paragraphs, or sentences negatively affects the agreement.
- *Readability measures*. We tested four standard readability measures, namely, *Ari* (Senter and Smith 1967), *Coleman-Liau* (Coleman and Liau 1975), *Flesch* (Flesch 1948), and *Lix* (Björnsson 1968) to identify whether readability of the documents plays any role in annotation agreement.

Correlation results are listed in Table 5. We observed the following statistically significant ($p < 0.05$) correlations. First, *document length* negatively correlates with agreement in *comments*. The longer the comment was the lower the agreement was. Second, *average paragraph length* negatively correlates with agreement in *blog posts*. The longer the paragraphs in blogs were, the lower agreement was reached. Third, all *readability scores* negatively correlate with agreement in the *public vs. private school* domain, meaning that the more complicated the text in terms of readability, the lower agreement was reached. We observed no significant correlation in *sentence coverage* and *average sentence length* measures. We cannot draw any general conclusion from these results, but we can state that some registers and topics, given their properties, are more challenging to annotate than others.

Probabilistic Confusion Matrix. Another qualitative analysis of disagreements between annotators was performed by constructing a *probabilistic confusion matrix* (Cinková, Holub, and Kríž 2012) on the token level.[20] The largest disagreements, as can be seen in Table 6, is caused by *rebuttal* and *refutation* confused with *none* (0.27 and 0.40, respectively). This is another sign that these two argument components were very hard to annotate. As shown in Table 4, the α_U was also low—0.08 for *rebuttal* and 0.17 for *refutation*.

20 "Properties: The sum in any row is 1. The *j*-th row of the matrix contains probabilities of assigning t_i given that another annotator has chosen t_j for the same instance. Thus, the *j*-th row of matrix describes expected tagging confusion related to the tag t_j." (Cinková, Holub, and Kríž 2012, page 846)

Table 5
Correlations between α_U and various measures on different data subsets. SC = full sentence coverage; DL = document length; APL = average paragraph length; ASL = average sentence length; ARI, C-L (Coleman-Liau), Flesch, LIX = readability measures. **Bold** numbers denote statistically significant correlation ($p < 0.05$).

	SC	DL	APL	ASL	ARI	C-L	Flesch	LIX
all data	−0.14	−0.14	0.01	0.04	0.07	0.08	−0.11	0.07
comments	−0.17	**−0.64**	0.13	0.01	0.01	0.01	−0.11	0.01
forum posts	−0.08	−0.03	−0.08	−0.03	0.08	0.24	−0.17	0.20
blog posts	−0.50	0.21	**−0.81**	−0.61	−0.39	0.47	0.04	−0.07
articles	0.00	−0.64	−0.43	−0.65	−0.25	0.39	−0.27	−0.07
homeschooling	−0.10	−0.29	−0.18	0.34	0.35	0.31	−0.38	0.46
redshirting	−0.16	0.07	−0.26	−0.07	0.02	0.14	−0.06	−0.09
prayer-in-school	−0.24	**−0.85**	0.30	0.07	0.14	0.11	−0.25	0.24
single sex	−0.08	−0.36	−0.28	−0.16	−0.17	0.05	0.06	0.06
mainstreaming	−0.27	−0.00	−0.03	0.06	0.20	0.29	−0.19	0.03
public private	0.18	0.19	0.30	−0.26	**−0.58**	**−0.51**	**0.51**	**−0.56**

Table 6
Probabilistic confusion matrix between all annotators.

	Claim	Premise	Backing	Rebuttal	Refutation	None
Claim	0.59	0.17	0.07	0.04	0.02	0.11
Premise	0.01	0.54	0.16	0.06	0.00	0.23
Backing	0.03	0.17	0.52	0.02	0.00	0.25
Rebuttal	0.16	0.12	0.10	0.32	0.03	0.27
Refutation	0.05	0.19	0.00	0.23	0.13	0.40
None	0.01	0.12	0.16	0.02	0.00	0.69

4.4.8 Causes of Disagreement—Qualitative Analysis and Problematic Phenomena. We analyzed the annotations and found the following phenomena that usually caused disagreements between annotators.

Granularity of Argument Components. Each argument component (e.g., *premise* or *backing*) should express one consistent and coherent piece of information, for example, a single reason in case of the *premise* (see Section 4.4.5). However, the decision whether a longer text should be kept as a single argument component or segmented into multiple components is subjective and highly text-specific.[21]

21 An example from Doc#4566 (artcomment, public-private-schools): One annotator labeled two premises: [*premise:* I send my kids to public schools because I care about them - their links with their diverse local community etc - but also because I care about the kind of culture they live in.] (summary: In public schools, kids have links with community and culture) [*premise:* To me, learning to care about and contribute to society as a whole - not just your own personal interests - is the best value a child can inherit.] (summary: At public, kids learn to contribute as a society). Another annotator labeled the same text as one single premise: [*premise:* I send my kids to public schools because I care about them - their links with their diverse local community etc - but also because I care about the kind of culture they live

Rhetorical Questions. Although rhetorical questions have been researched extensively in linguistics (Schmidt-Radefeldt 1977; Han 2002; Egg 2007; Lee-Goldman 2006), their role in argumentation represents a substantial research question (Petty, Cacioppo, and Heesacker 1981; Frank 1990; Roberts and Kreuz 1994; Ilie 1999; Ottati, Rhoads, and Graesser 1999). Teninbaum (2011) provides a brief history of rhetorical questions in persuasion. In short, rhetorical questions should provoke the reader. From the perspective of our argumentation model, rhetorical questions might fall both into the *logos* dimension (and thus be labeled as *claim*, *premise*, etc.) or into the *pathos* dimension (refer to Section 2.2). Again, the decision is usually not clear-cut.

Refutation versus Premise. As introduced in Section 4.4.2, *rebuttal* attacks the *claim* by presenting an opponent's view. In most cases, the *rebuttal* is again attacked by the author using *refutation*. From the pragmatic perspective, *refutation* thus supports the author's stance expressed by the *claim*. Therefore, it can be easily confused with *premises*, as the function of both is to provide support for the *claim*. *Refutation* thus only takes place if it is meant as a reaction to the *rebuttal*. It follows the discussed matter and contradicts it. Such a discourse is usually expressed as:

> [*claim:* My claim.] [*rebuttal:* On the other hand, some people claim XXX which makes my claim wrong.] [*refutation:* But this is not true, because of YYY.]

However, the author might also take the following defensible approach to formulate the argument:

> [*rebuttal:* Some people claim XXX-1 which makes my claim wrong.] [*refutation:* But this is not true, because of YYY-1.] [*rebuttal:* Some people claim XXX-2 which makes my claim wrong.] [*refutation:* But this is not true, because of YYY-2.] [*claim:* Therefore my claim.]

If this argument is formulated without stating the *rebuttals*, it would be equivalent to the following:

> [*premise:* YYY-1.] [*premise:* YYY-2.] [*claim:* Therefore my claim.]

This example shows that *rebuttal* and *refutation* represent a rhetorical device to produce arguments, but the distinction between *refutation* and *premise* is context-dependent and on the functional level both *premise* and *refutation* have very similar roles—to support the author's standpoint. Although introducing dialogical moves into a monological model and its practical consequences, as described above, can be seen as a shortcoming of our model, this rhetoric figure has been identified by argumentation researchers as *procatalepsis* (Walton 2007, page 106). A broader view on incorporating opposing views (or lack thereof) is discussed under the term *confirmation bias* by Mercier and Sperber (2011, page 63), who claim that "people are trying to convince others. They are typically looking for arguments and evidence to confirm their own claim, and ignoring negative arguments and evidence unless they anticipate having to rebut them." The dialectical attack of possible counter-arguments may thus strengthen one's own argument.

in. To me, learning to care about and contribute to society as a whole - not just your own personal interests - is the best value a child can inherit.] (summary: Kids should learn the cultural diversity).

One possible solution would be to refrain from capturing this phenomenon completely and to simplify the model to claims and premises, for instance. However, the following example would then miss an important piece of information, as the last two clauses would be left un-annotated. At the same time, annotating the last clause as *premise* would be misleading, because it does not support the *claim.* More precisely, it supports the claim only indirectly by attacking the *rebuttal*. (We consider here that a support is an admissible extension of an abstract argument graph, as suggested by Dung [1995]).

> **Doc#422 (forumpost, homeschooling)** [*claim:* I try not to be anti-homeschooling, but... it's just hard for me.] [*premise:* I really haven't met any homeschoolers who turned out quite right, including myself.] I apologize if what I'm saying offends any of you - that's not my intention, [*rebuttal:* I know that there are many homeschooled children who do just fine,] but [*refutation:* that hasn't been my experience.]

To the best of our knowledge, these context-dependent dialogical properties of argument components using Toulmin's model have not been solved in the literature on argumentation theory and we suggest that these observations should be taken into account in the future research in monological argumentation.

Purely Sarcastic Argumentation and Fallacies in General. Appeal to emotion, sarcasm, irony, or jokes are common in argumentation in user-generated Web content. We also observed documents in our data that were purely sarcastic (the *pathos* dimension); therefore logical analysis of the argument (the *logos* dimension) would make no sense. However, given the structure of such documents, some *claims* or *premises* might also be identified. Such an argument is a typical example of fallacious argumentation, which intentionally *pretends* to present a valid argument; but its persuasion is conveyed purely, for example, by appealing to emotions of the reader (Tindale 2007).

4.4.9 Analysis of Annotated Corpus from the Argumentation Research Perspective. We present some statistics of the annotated data that are important from the argumentation research perspective. Regardless of the register, 48% of *claims* are implicit. This means that the authors assume that their standpoint towards the discussed controversy can be inferred by the reader and give only reasons for that standpoint. Also, explicit *claims* are mainly written just once; the *claim* was rephrased and occurred multiple times in only 3% of the documents.

In 6% of the documents, the reasons for an implicit *claim* are given only in the *pathos* dimension, making the argument purely persuasive without logical argumentation.

The "myside bias," defined as a bias against information supporting another side of an argument (Perkins 1985; Wolfe, Britt, and Butler 2009), can be observed by the presence of *rebuttals* to the author's *claim* or by formulating arguments for both sides when the overall stance is neutral. Thus, 85% of the documents do not consider any opposing side, and only 8% of the documents present a *rebuttal*, which is then attacked by *refutation* in 4% of the documents. Multiple *rebuttals* and *refutations* were found in 3% of the documents. Only 4% of the documents were overall neutral and presented arguments for both sides, mainly in blog posts.

Hedging in Claims. We were also interested in whether mitigating linguistic devices are used in the annotated arguments, namely, in their main stance-taking components, the *claims*. Such devices typically include parenthetical verbs, syntactic constructions, token agreements, hedges, challenge questions, discourse markers, and tag questions,

among others (Flores-Ferrán and Lovejoy 2015). In particular, Kaltenbök, Mihatsch, and Schneider (2010, page 1) define hedging as "a discourse strategy that reduces the force or truth of an utterance and thus reduces the risk a speaker runs when uttering a strong or firm assertion or other speech act." We manually examined the use of hedging in the annotated *claims*.

Our main observation is that hedging is used differently across topics. For instance, about 30% to 35% of claims in *homeschooling* and *mainstreaming* signal the lack of a full commitment to the expressed stance, in contrast to *prayer in schools* (15%) or *public vs. private schools* (about 10%). Typical hedging cues include speculations and modality (***If** I have kids, I will **probably** homeschool them.*), statements as neutral observations (***It's not wrong to hold the opinion that in general it's better** for kids to go to school than to be homeschooled.*), or weasel[22] phrases (Farkas et al. 2010) (***In some cases**, inclusion can work fantastically well., **For the majority of the children in the school**, mainstream would not have been a suitable placement.*).

On the other hand, most *claims* that are used for instance in the *prayer in schools* arguments are very direct, without trying to diminish its commitment to the conveyed belief (for example, *NO PRAYER IN SCHOOLS!... period.*, *Get it out of public schools*, *Pray at home.*, or *No organized prayers or services anywhere on public school board property - FOR ANYONE.*). Moreover, some claims are clearly offensive, persuading by direct imperative clauses towards the opponents/audience (*TAKE YOUR KIDS PRIVATE IF YOU CARE AS I DID*, *Run, don't walk, to the nearest private school.*) or even accuse the opponents for taking a certain stance (*You are a bad person if you send your children to private school.*).

These observations are consistent with the findings from the first annotation study on persuasion (see Section 4.3.2), namely, that some topics attract *heated* argumentation, where participants, take very clear and reserved standpoints (such as *prayer in schools* or *private vs. public schools*), whereas discussions about other topics are rather milder. It has been shown that the choices a speaker makes to express a position are informed by their social and cultural background, as well as their ability to speak the language (Dippold 2007; Kreutel 2007; Flores-Ferrán and Lovejoy 2015). However, given the uncontrolled settings of the user-generated Web content, we cannot infer any similar conclusions in this respect.

Analyzing Type of Support. We investigated *premises* across all topics in order to find the type of support used in the argument. We followed the approach of Park and Cardie (2014), who distinguished three types of propositions in their study, namely, *unverifiable*, *verifiable non-experiential*, and *verifiable experiential*.

Verifiable non-experiential and *verifiable experiential* propositions, unlike *unverifiable propositions*, contain an objective assertion, where objective means "expressing or dealing with facts or conditions as perceived without distortion by personal feelings, prejudices, or interpretations."[23] Such assertions have truth values that can be proved or disproved with objective evidence; the correctness of the assertion or the availability of the objective evidence does not matter (Park and Cardie 2014, page 31). A verifiable proposition can further be distinguished as experiential or not, depending on whether the proposition is about the writer's personal state or experience or something non-experiential. Verifiable experiential propositions are sometimes referred to as anecdotal

22 https://en.wikipedia.org/wiki/Weasel_word.
23 http://www.merriam-webster.com/dictionary/objective.

evidence, and provide the novel knowledge that readers are seeking (Park and Cardie 2014, page 31).

Table 7 shows the distribution of the premise types with examples for each topic from the annotated corpus. As can be seen in the first row, arguments in *prayer in schools* contain a majority (73%) of unverifiable premises. Closer examination reveals that their content varies from general vague propositions to obvious fallacies, such as a hasty generalization, straw men, or slippery slope. As Nieminen and Mustonen (2014) identified, fallacies are very common in argumentation about religion-related issues. On the other side of the spectrum, arguments about *redshirting* rely mostly on anecdotal evidence (61% of *verifiable experiential* propositions). We will discuss the phenomena of narratives in argumentation in more detail later in Section 4.4.10. All the topics except *private vs. public schools* exhibit similar amounts of *verifiable non-experiential* premises (9% to 22%), usually referring to expert studies or facts. However, this type of premise has usually the lowest frequency.

4.4.10 Discussion. Manually analyzing argumentative discourse and reconstructing (annotating) the underlying argument structure and its components is difficult. As (Reed and Rowe 2006, page 267) point out, "the analysis of arguments is often hard, not only for students, but for experts too." According to Harrell (2011a, page 81), argumentation is a skill and "even for simple arguments, untrained college students can identify the conclusion but without prompting are poor at both identifying the premises and how the premises support the conclusion." Harrell (2011b, page 81) further claims that "a wide literature supports the contention that the particular skills of understanding, evaluating, and producing arguments are generally poor in the population of people who have not had specific training and that specific training is what improves these skills." Some studies, for example, show that students perform significantly better on reasoning tasks when they have learned to identify premises and conclusions (Shaw 1996) or have learned some standard argumentation norms (Weinstock, Neuman, and Tabak 2004).

One particular extra challenge in analyzing argumentation in Web user-generated discourse is that the authors produce their texts probably without any existing argumentation theory or model in mind.[24] We assume that argumentation or persuasion is inherent when users discuss controversial topics, but the true reasons why people participate in on-line communities and what drives their behavior is another research question (Bishop 2007; Cullen and Morse 2011; de Melo Bezerra and Hirata 2011; Sun, Rau, and Ma 2014). When the analyzed texts have a clear intention to produce argumentative discourse, such as in argumentative essays (Stab and Gurevych 2014a), the argumentation is much more explicit and substantially higher inter-annotator agreement can be achieved.

Suitability of the Modified Toulmin Model. The model seems to be suitable for short persuasive documents, such as comments and forum posts. Its applicability to longer documents, such as articles or blog posts, is problematic for several reasons.

The argument components of the (modified) Toulmin model and their roles are not expressive enough to capture argumentation that not only conveys the logical

24 By analyzing the arguments during annotation, our impression was that an average user participating in online discussions is not a skilled arguer. However, we lack grounded empirical evidence to support such a claim.

Table 7
Distribution of premise types for each topic with examples. HS = homeschooling; MS = mainstreaming; PIS = prayer in schools; PPS = private vs. public schools; RS = redshirting; SSE = single sex education. Number of analyzed premises shown in parentheses.

	Unverifiable	Verifiable non-experiential	Verifiable experiential
PIS (112)	73% • *Religion is basically a gang mentality where people feel they need to belong to a group...* • *A primary purpose of public education is to shape good citizens.*	22% • *Fact: Muslims pray five times daily, in a way which is not practical in a normal classroom setting.* • *Japan, where no one prays at school, has the lowest crime rate of any developed nation.*	4% • *When I was a kid we learned religion in church, math, reading, history, etc., in school and at home.* • *I am a victim of this latter possibility. Believe me, I'm still trying to repair the damage.*
HS (160)	57% • *But when you put 30 kids in one classroom, it becomes very difficult to teach them all individually.* • *The trouble is, home schooling can be a cover for all sorts of undesirable stuff.*	14% • *Only a fortnight ago a report was published by Robin Alexander and his team at Cambridge University which found that the primary school curriculum is too narrow and involves too much testing.* • *Dr. Smedley believes that homeschoolers have superior socialization skills, and his research supports this claim.*	29% • *It was boring, tedious, slow and frustrating. I learned nothing that I did not know before other than a handful of French verbs which have so far been of as much use as a chocolate fireguard.* • *Everyone I know went to public school and on to college. We didn't feel unprepared*
SSE (96)	46% • *Co-ed schools don't cultivate the cooperation or better understanding between the opposite sexes.* • *Research is quite clear about this.*	18% • *Studies show that women suffer from a stereotype threat in math and science, meaning that in the fields of math and science women are more apprehensive to perform [...]* • *Studies clearly establish that single-sex schools are, IN GENERAL, better for educational outcomes than co-ed schools.*	36% • *The unhealthiest social situations I was ever in were an all-boys school and military training.* • *Once I switched schools, I found myself with valuable ? and what I?m sure to be life long ? friendships with the boys I sat with in class.*
MS (51)	47% • *School and classroom design has not evolved and this has stagnated the inclusion movement.* • *The level of differentiated instruction required to develop some functional skills is not possible in mainstream classrooms.*	10% • *In a TEACH magazine article about his new book, Adelman says inclusion of students with disabilities benefits entire student bodies by teaching kids about diversity [...]* • *The reality is students with special needs are a small percentage of the population and cannot drive a fundamental shift in education.*	43% • *I have a HF autistic son w/ severe ADHD.. He is doing awesome in grade 1 he has a 1 on 1 aide in the class.. We feel well supported by the school system and he has only 18 in his class!* • *I often spent 20 mins of each year 9 lesson getting the boys to stop aggravating the ADHD boy, as he would then "blow", much to the amusement of everyone.*
PPS (159)	43% • *Public schools are not about education; they are about social engineering.* • *Kids are indoctrinated not educated in Public Schools.*	0% —	57% • *Worked for us.* • *I have five Daughters and they all went to private schools and everyone of them have a degree and now have good paying jobs.*
RS (67)	30% • *they will grow up,, they will mature.* • *These kids need to be prepared for the 21st centuary global economy by being enrolled in a local second language immersion kindergarden as soon as they can enroll.*	9% • *There have been a lot of studies that show people born later in the year (March) are more successful in life due to the age gap.* • *Studies show the practice (when common) has socioeconomic repurcissions down the line and can increase the HS drop out rate [...]*	61% • *I honestly made my decision because I had a choice and I just did not feel right personally about sending my DS to Kindergarten at age 4.* • *However with my oldest if he were born a couple of months earlier I would have kept him back a year as he would not have been ready.*

structure (in terms of reasons put forward to support the claim), but also relies heavily on the rhetorical power. This involves various stylistic devices, pervading narratives, direct and indirect speech, or interviews.[25] Although in some cases the argument components are easily recognizable, the vast majority of the discourse in articles and blog posts does not correspond to any distinguishable argumentative function in the *logos* dimension. As the purpose of such discourse relates more to rhetoric than to argumentation, unambiguous analysis of such phenomena goes beyond capabilities of the current argumentation model. For a discussion about metaphors in Toulmin's model of argumentation see, for example, Xu and Wu (2014), Santibáñez (2010).

Articles without a clear standpoint towards the discussed controversy cannot be easily annotated with the model either. Although the matter is viewed from both sides and there may be reasons presented for either of them, the overall persuasive intention is missing and fitting such data to the argumentation framework causes disagreements.[26] One solution might be to break the document down to paragraphs and annotate each paragraph separately, examining argumentation on a different level of granularity.

Annotating Other Dimensions of Argument. As introduced in Section 2.2, there are several dimensions of an argument. The Toulmin model focuses solely on the *logos* dimension. We decided to ignore the *ethos* dimension, because dealing with an author's credibility remains unclear, given the variety of the source Web data.[27] However, exploiting the *pathos* dimension of an argument is prevalent in the Web data, for example, as an appeal to emotions. Therefore we experimented with annotating *appeal to emotions* as a separate category independent of components in the *logos* dimension. We defined some features for the annotators on how to distinguish *appeal to emotions*. Figurative language such as hyperbole, sarcasm, or obvious exaggerating to "spice up" the argument are the typical signs of pathos. In an extreme case, the whole argument might be purely emotional, as in the following example.

> **Doc#1698 (comment, prayer in schools)** [*app-to-emot:* Prayer being removed from school is just the leading indicator of a nation that is 'Falling Away' from Jehovah. [...] And the disasters we see today are simply God's finger writing on the wall: Mene, mene, Tekel, Upharsin; that is, God has weighed America in the balances, and we've been found wanting. No wonder 50 million babies have been aborted since 1973. [...]]

We kept annotations on the *pathos* dimension as simple as possible (with only one *appeal to emotions* label), but the resulting agreement was unsatisfying ($\alpha_U = 0.30$) even after several annotation iterations. Appeal to emotions is considered a type of fallacy (Govier 2010; Damer 2013). Given the results, we assume that a more carefully designed approach to fallacy annotation should be applied. To the best of our knowledge, there has been very little research on modeling fallacies similarly to arguments at the discourse level (Pineau 2013). Therefore the question, in which detail and structure fallacies should be annotated, remains open. For the rest of the article, we thus focus on the *logos* dimension solely.

25 For a deep analysis of the role of direct speech in newspaper discourse argumentation, see Smirnova (2009).

26 Note that we only filtered persuasive documents in annotation study 1 (Section 4.3) for comments and forum posts; blog posts and newswire articles were checked only briefly while collecting the *raw corpus*.

27 Modeling influential persons belongs to research in social network analysis, which is beyond the scope of this article.

Narratives in Argumentation. Some of the educational topics under examination relate to young children (e.g., redshirting or mainstreaming); therefore we assume that the majority of participants in discussions are their parents. We observed that many documents related to these topics contain narratives. Sometimes the storytelling is meant as a support for the argument, but there are documents where the narrative has no intention of persuading and is simply a story sharing.

There is no widely accepted theory of the role of narratives among argumentation scholars. According to Fisher (1987), humans are storytellers by nature, and the "reason" in argumentation is therefore better understood in and through the narratives. He found that good reasons often take the form of narratives. Hoeken and Fikkers (2014) investigated how integration of explicit argumentative content into narratives influences issue-relevant thinking and concluded that identifying with the character who is in favor of the issue yielded a more positive attitude toward the issue. In recent research, Bex (2011) proposes an argumentative-narrative model of reasoning with evidence, further elaborated in Bex, Bench-capon, and Verheij (2012); in addition, Niehaus et al. (2012) proposes a computational model of narrative persuasion.

Stemming from another research field, Leyton Escobar, Kommers, and Beldad (2014) found that online community members who use and share narratives have higher participation levels and that narratives are useful tools to build cohesive cultures and increase participation. Betsch et al. (2010) examined influencing vaccine intentions among parents and found that narratives carry more weight than statistics.

4.5 Summary of Annotation Studies

This section described two annotation studies that deal with argumentation in user-generated Web content on different levels of detail. In Section 4.3, we argued for a need for document-level distinction of persuasiveness. We annotated 990 comments and forum posts, reaching moderate inter-annotator agreement (Fleiss' $\pi = 0.59$). Section 4.4 motivated the selection of a model for micro-level argument annotation, proposed its extension based on pre-study observations, and outlined the annotation set-up. This annotation study resulted into 340 documents annotated with the modified Toulmin model and reached moderate inter-annotator agreement in the logos dimension (Krippendorff's $\alpha_U = 0.48$). These results make the annotated corpora suitable for training and evaluation computational models, and each of these two annotation studies will have their experimental counterparts in the following section.

5. Experiments

This section presents experiments conducted on the annotated corpora introduced in Section 4. We put the main focus on identifying *argument components* in the discourse.[28] To comply with machine learning terminology, in this section we will use the term

28 We also experimented with classification of *persuasive* documents, as introduced in Annotation Study 1 (section 4.3). This task can be seen as standard document-level two-class text classification. Using SVM (Cortes and Vapnik 1995) with Sequential Minimal Optimization (Platt 1999), polynomial kernel, and n-gram baseline features, we obtained 0.69 Macro-F_1 score. We also used a rich feature set (a large part of features that will be discussed in Section 5.1) but the system did not beat the baseline, therefore we do not report on this experiment in detail. However, we expect that in a real-world scenario of automatically analyzing argument components in user-generated content, the first step of assessing on-topic persuasiveness (or external relevance [Paglieri and Castelfranchi 2014]) is essential.

domain as an equivalent to a topic (remember that our data set includes six different topics; see Section 4.1).

We evaluate three different scenarios. First, we report *ten-fold cross-validation* over a random ordering of the entire data set. Second, we deal with *in-domain ten-fold cross-validation* for each of the six domains. Third, in order to evaluate the domain portability of our approach, we train the system on five domains and test on the remaining one for all six domains (which we report as *cross-domain validation*).

5.1 Identification of Argument Components

In the following experiment, we focus on automatic identification of arguments in the discourse. Our approach is based on supervised and semi-supervised machine learning methods on the *gold data Toulmin* data set introduced in Section 4.4.

An argument consists of different components (such as *premises*, *backing*, etc.) which are implicitly linked to the *claim*. In principle, one document can contain multiple independent arguments.[29] However, only 4% of the documents in our data set contain arguments for both sides of the issue. Thus, we simplify the task and assume there is only one argument per document.[30]

Given the low inter-annotator agreement on the *pathos* dimension (Table 4), we focus solely on recognizing the logical dimension of argument. The *pathos* dimension of argument remains an open problem for proper modeling as well as its later recognition.

5.1.1 Data Representation and Evaluation. Because the smallest annotation unit is a token and the argument components do not overlap, we approach identification of argument components as a sequence labeling problem. We use the BIO encoding, so each token belongs to one of the following 11 classes: *O* (not a part of any argument component), *Backing-B*, *Backing-I*, *Claim-B*, *Claim-I*, *Premise-B*, *Premise-I*, *Rebuttal-B*, *Rebuttal-I*, *Refutation-B*, *Refutation-I*. This is the minimal encoding that is able to distinguish two adjacent argument components of the same type. In our data, 48% of all adjacent argument components of the same type are direct neighbors (there are no "O" tokens in between).

We report Macro-F_1 score and F_1 scores for each of the 11 classes as the main evaluation metric. This evaluation is performed on the token level, and for each token the predicted label must exactly match the gold data label (classification of tokens into 11 classes).

As instances for the sequence-labeling model, we chose *sentences* rather than *tokens*. During our initial experiments, we observed that building a sequence labeling model for recognizing argument components as sequences of *tokens* is too fine-grained, as a

29 In our approach to annotation of controversies, this would mean that the *overall* standpoint of the author is neutral but she presents arguments for both sides of the controversy.

30 This simplification can be seen as a limitation of our model, as argumentation mining in some related works is a form of structured predictions of elements in discourse where the explicit notion of relation between argument components is crucial for argument "parsing," for example, in the work by Peldszus and Stede (2015) envisioned in their earlier survey paper (Peldszus and Stede 2013a), or by Stab and Gurevych (2014b). It is thus possible that in a general argumentative discourse, the same proposition can play two different roles in two arguments, similarly to the approach of Aharoni et al. (2014). This phenomena was discussed as *divergent structures* by Thomas (1981) and later elaborated on by Freeman (2011, page 16).

single token does not convey enough information that could be encoded as features for a machine learner. However, as discussed in Section 4.4.5, the annotations were performed on data pre-segmented to sentences and annotating tokens was necessary only when the sentence segmentation was wrong or one sentence contained multiple argument components. Our corpus consists of 3,899 sentences, from which 2,214 sentences (57%) contain no argument components. From the remaining ones, only 50 sentences (1%) have more than one argument component. Although in 19 cases (0.5%) the sentence contains a *Claim–Premise* pair, which is an important distinction from the argumentation perspective, given the overall small number of such occurrences, we simplify the task by treating each sentence as if it has either one argument component or none.

The approximation with sentence-level units is explained in the example in Figure 8. In order to evaluate the expected performance loss using this approximation, we used an *oracle* that always predicts the correct label for the unit (sentence) and evaluated it against the true labels (recall that the evaluation against the true gold labels is done always on token level). We lose only about 10% of Macro-F_1 score (0.906) and only about 2% of accuracy (0.984). This performance is still acceptable, while allowing us to model sequences where the minimal unit is a sentence.

5.1.2 Gold Data Statistics. Table 8 shows the distribution of the classes in the *gold data Toulmin*, where the labeling was already mapped to the sentences. The minimal presence of *rebuttal* and *refutation* (four classes account for only 3.4% of the data) makes this data set very unbalanced.

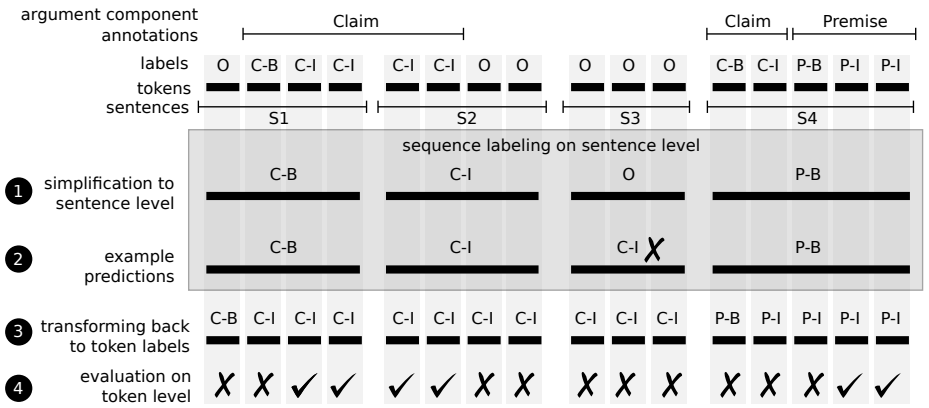

Figure 8
Our approach to simplifying argument component segmentation and evaluation of the system. Gold data are labeled on the token level (*C* = *Claim*, *P* = *Premise*). In step 1, the argument component label becomes a new label for the entire sentence. The resulting label reflects if the component begins in the sentence (i.e., the case of *Claim-B* in *S1*). If there are more components in one sentence, the longest one is selected (i.e., the case of *Claim* and *Premise* in *S4*). In step 2, the predictions are obtained for entire sentences, as one sentence represents the minimal unit in sequence labeling mode. In step 3, the labels are translated back to the token level, spanning over the entire sentence. If the predicted label is a **-B* tag, the first token is labeled as **-B* and the remaining ones as **-I* (i.e., in *S1* and *S4*). In step 4, evaluation of predictions is performed solely on the token level by comparing the predicted token labels with the gold token labels.

Table 8
Class distribution of the *gold data Toulmin* corpus approximated to the sentence-level boundaries.

Class	Sentences in data Relative (%)	Absolute	Class	Sentences in data Relative (%)	Absolute
Backing-B	5.6	220	Premise-I	8.6	336
Backing-I	7.2	281	Rebuttal-B	1.6	61
Claim-B	4.4	171	Rebuttal-I	0.9	37
Claim-I	0.4	16	Refutation-B	0.5	18
O	56.8	2214	Refutation-I	0.4	15
Premise-B	13.6	530	**Total**		3899

5.1.3 Methods and Features. We chose SVM^{hmm} (Joachims, Finley, and Yu 2009) implementation[31] of Structural Support Vector Machines[32] for sequence labeling.[33] Each sentence ($\mathbf{x}$) is represented as a vector of real-valued features.

We defined the following feature sets:

- **FS0**: Baseline lexical features
  - word uni-, bi-, and tri-grams (binary)
- **FS1**: Structural, morphological, and syntactic features
  - First and last three tokens. *Motivation:* these tokens may contain discourse markers or other indicators for argument components, such as *therefore* and *since* for premises or *think* and *believe* for claims.
  - Relative position in paragraph and relative position in document. *Motivation:* We expect that claims are more likely to appear at the beginning or at the end of the document.
  - Number of POS 1-3 grams, dependency tree depth, constituency tree production rules, and number of subclauses. Based on Stab and Gurevych (2014b).
- **FS2**: Topic and sentiment features
  - 30 features taken from a vector representation of the sentence obtained by using Gibbs sampling on LDA model (McCallum 2002; Blei, Ng, and Jordan 2003) with 30 topics trained on unlabeled data from the *raw corpus*. *Motivation:* Topic representation of a sentence might be valuable for detecting off-topic sentences, namely, non-argument components.

31 `http://www.cs.cornell.edu/people/tj/svm_light/svm_hmm.html`.

32 Another widely used method for sequence labeling is CRF (Lafferty, McCallum, and Pereira 2001), but the performance of CRF has been found comparable to SVM^{hmm} by Keerthi and Sundararajan (2007).

33 Argument components can span several sentences, their boundaries are not fixed. Therefore, each sentence belonging to a particular argument component can be encoded with two different tags, namely, the begin tag (i.e., *Claim-B*) for the first sentence and the "in" tag (i.e., *Claim-I*) for the following sentences. Although this can be treated as a simple sentence classification task, sequence labeling can leverage the probability distribution of label sequences (for instance *Claim-B*, *Claim-I* are more likely to occur than *Claim-B*, *Premise-I*).

  - Scores for five sentiment categories (from very negative to very positive) obtained from the Stanford sentiment analyzer (Socher et al. 2013). *Motivation:* Claims usually express opinions and carry sentiment.

- **FS3**: Semantic, coreference, and discourse features
  - Binary features from Clear NLP Semantic Role Labeler (Choi 2012). Namely, we extract *agent*, *predicate + agent*, *predicate + agent + patient + (optional) negation*, *argument type + argument value*, and *discourse marker*, which are based on PropBank semantic role labels.[34] *Motivation:* Exploit the semantics of capturing the semantics of the sentences.
  - Binary features from Stanford Coreference Chain Resolver (Lee et al. 2013), for example, presence of the sentence in a chain, transition type (i.e., nominal–pronominal), distance to previous/next sentences in the chain, or number of inter-sentence coreference links. *Motivation:* Presence of coreference chains indicates links outside the sentence and thus may be informative, for example, for classifying whether the sentence is a part of a larger argument component.
  - Results of a PTDB-style discourse parser (Lin, Ng, and Kan 2014), namely, the type of discourse relation (explicit, implicit), presence of discourse connectives, and attributions. *Motivation:* It has been claimed that discourse relations play a role in argumentation mining (Cabrio, Tonelli, and Villata 2013).

- **FS4**: Embedding features
  - 300 features from word embedding vectors using word embeddings trained on part of the Google News data set (Mikolov et al. 2013). In particular, we sum up embedding vectors (dimensionality 300) of each word, resulting in a single vector for the entire sentence. This vector is then directly used as a feature vector.[35] *Motivation:* Embeddings helped to achieve state-of-the-art results in various NLP tasks (Socher et al. 2013; Guo et al. 2014).

Except for the baseline lexical features, all feature types are extracted not only for the current sentence s_i, but also for C preceding and subsequent sentences, namely, s_{i-C}, s_{i-C+1}, ... s_{i+C-1}, s_{i+C}, where C was empirically set to 4.[36] Each feature is then represented with a prefix to determine its relative position to the current sequence unit.[37]

5.1.4 Results. Let us first discuss the upper bounds of the system. Performance of the three human annotators is shown in the first column of Table 9 (results are obtained

34 Explained in detail in annotation guidelines at `http://clear.colorado.edu/compsem/documents/propbank_guidelines.pdf`.

35 Le and Mikolov (2014) proposed more advanced techniques for sentence representation using embeddings.

36 We used grid search with different C values in several feature set combinations (FS01, FS012) over the entire cross-validation scenario and fixed the value afterwards.

37 For example, `minus2Sent_sentimentNegative=0.23` or `plus1Sent_DependencyTreeDepth=3`.

Table 9
Ten-fold cross validation results of classification of argument components using different feature sets. Macro-F_1 and F_1 scores for individual classes are shown. Column Hum denotes human performance, Ran is a random classifier, "O" is a majority voting (each token is labeled as "O"). **Bold** numbers denote the best results for the given class. The best performing configuration is the 234 feature set. Differences between the best feature sets (01234 and 234) and other sets are statistically significant ($p < 0.001$, paired exact Liddell's test).

				Feature set combinations								
	Hum	**Ran**	**"O"**	0	01	012	0123	01234	1234	234	34	4
M-F_1	.602	.071	.065	.156	.217	.229	.219	**.251**	.240	**.251**	.238	.229
Bac-B	.664	.063	.000	.140	.262	.311	.294	.320	**.326**	.291	.278	.278
Bac-I	.579	.120	.000	.159	.339	.364	.334	.372	**.380**	.366	.362	.363
Cla-B	.739	.051	.000	.127	.203	.234	.211	.257	.259	**.270**	.266	.252
Cla-I	.728	.051	.000	.165	.207	.224	.194	.237	.245	**.269**	.258	.242
O	.833	.136	.707	**.714**	.705	.703	.708	.691	.686	.675	.671	.669
Pre-B	.673	.082	.000	.176	.280	.289	.286	**.298**	.294	.265	.269	.246
Pre-I	.736	.179	.000	.241	.390	.391	.380	**.400**	.396	.357	.366	.356
Reb-B	.403	.026	.000	.000	.000	.000	.000	.082	.000	**.037**	.000	.035
Reb-I	.495	.053	.000	.000	.000	.000	.000	.104	.054	**.118**	.036	.076
Ref-B	.390	.000	.000	.000	.000	.000	.000	.000	.000	**.057**	**.057**	.000
Ref-I	.387	.023	.000	.000	.000	.000	.000	.000	.000	**.055**	.052	.000

from a cumulative confusion matrix). The overall Macro-F_1 score is 0.602 (accuracy = 0.754). If we look closer at the different argument components, we observe that humans are good at predicting *claims*, *premises*, *backing*, and non-argumentative text (F_1 about 0.60–0.80), but on *rebuttal* and *refutation* they achieve rather low scores. Without these two components, the overall human Macro-F_1 would be 0.707. This trend follows the inter-annotator agreement scores, as discussed in Section 4.4.6.

In our experiments, the feature sets were combined in the bottom–up manner, starting with the simple lexical features (FS0), adding structural and syntactic features (FS1), then adding topic and sentiment features (FS2), then features reflecting the discourse structure (FS3), and finally enriched with completely unsupervised latent vector space representation (FS4). In addition, we were gradually removing the simple features (without lexical features, without syntactic features, etc.) to test the system with more "abstract" feature sets (feature ablation). The results are shown in Table 9.

The overall best performance (Macro-$F_1 = 0.251$) was achieved using the rich feature sets (01234 and 234) and significantly outperformed the baseline as well as other feature sets. Classification of non-argumentative text (the "O" class) yields an F_1 score of about 0.7 even in the baseline setting. The boundaries of *claims* (Cla-B), *premises* (Pre-B), and *backing*, (Bac-B) reach on average lower scores than their respective inside tags (Cla-I, Pre-I, Bac-I). It can be interpreted such that the system is able to classify that a certain sentence belongs to a certain argument component, but the distinction whether it is a beginning of the argument component is harder. The very low numbers for *rebuttal* and *refutation* have two reasons. First, these two argument components caused many disagreements in the annotations, as discussed in Section 4.4.8, and were hard to recognize for the humans, too. Second, these four classes have very few instances in the corpus (about 3.4%, see Table 8), so the classifier suffers from the lack of training data.

The results for the *in-domain cross validation scenario* are shown in Table 10. Similarly to the cross-validation scenario, the overall best results were achieved using the largest feature set (01234). For *mainstreaming* and *red-shirting*, the best results were achieved using only the feature set 4 (embeddings). These two domains also contain fewer documents, compared with other domains (refer to Table 2). We suspect that embeddings-based features convey important information when not enough in-domain data are available. This observation will become apparent in the next experiment.

The *cross-domain* experiments yield rather poor results for most of the feature combinations (Table 11). However, using only feature set 4 (embeddings), the system performance increases rapidly, so it is even comparable to numbers achieved in the *in-*

Table 10
Results of classification of argument components in the in-domain cross-validation scenario. Macro-F_1 scores reported, **bold** numbers denote the best results. HS = homeschooling, MS = mainstreaming; PIS = prayer in schools; PPS = private vs. public schools; RS = redshirting; SSE = single sex education. Results in the *aggregated* row are computed from an aggregated confusion matrix over all domains. The differences between the best feature set combination (01234) and others are statistically significant ($p < 0.001$; paired exact Liddell's test).

	Feature set combinations								
Domain	0	01	012	0123	01234	1234	234	34	4
HS	0.134	0.162	0.167	0.165	0.187	0.176	**0.205**	0.203	0.193
MS	0.072	0.123	0.138	0.151	0.198	0.216	0.165	0.190	**0.226**
PIS	0.152	0.174	0.178	0.168	**0.212**	0.192	0.175	0.177	0.181
PPS	0.235	0.233	0.230	0.240	**0.265**	0.250	0.239	0.250	0.243
RS	0.090	0.156	0.156	0.144	0.195	0.201	0.204	0.190	**0.225**
SSE	0.141	0.176	0.200	0.185	0.206	**0.216**	0.189	0.202	0.201
Aggregated	0.182	0.200	0.205	0.206	**0.236**	0.230	0.218	0.228	0.229

Table 11
Results of classification of argument components in the cross-domain scenario. Macro-F_1 scores reported, **bold** numbers denote the best results. HS = homeschooling; MS = mainstreaming; PIS = prayer in schools; PPS = private vs. public schools; RS = redshirting; SSE = single sex education. Results in the *aggregated* row are computed from an aggregated confusion matrix over all domains. The differences between the best feature set combination (4) and others are statistically significant ($p < 0.001$; paired exact Liddell's test).

	Feature set combinations								
Domain	0	01	012	0123	01234	1234	234	34	4
HS	0.087	0.063	0.044	0.106	0.072	0.075	0.065	0.063	**0.197**
MS	0.072	0.060	0.070	0.058	0.038	0.062	0.045	0.060	**0.188**
PIS	0.078	0.073	0.083	0.074	0.086	0.073	0.096	0.081	**0.166**
PPS	0.070	0.059	0.070	0.132	0.059	0.062	0.071	0.067	**0.203**
RS	0.067	0.067	0.082	0.110	0.097	0.092	0.075	0.075	**0.257**
SSE	0.092	0.089	0.066	0.036	0.120	0.091	0.071	0.066	**0.194**
Aggregated	0.079	0.086	0.072	0.122	0.094	0.088	0.089	0.076	**0.209**

domain scenario. These results indicate that embedding features generalize well across domains in our task of argument component identification. We leave investigating better performing vector representations, such as *paragraph vectors* (Le and Mikolov 2014), for future work.

5.1.5 Error Analysis. Error analysis based on the probabilistic confusion matrix (Wang et al. 2013) shown in Table 12 reveals further details. About a half of the instances for each class are misclassified as non-argumentative (the "O" prediction).

Backing-B is often confused with *Premise-B* (12%) and *Backing-I* with *Premise-I* (23%). Similarly, *Premise-I* is misclassified as *Backing-I* in 9%. This shows that distinguishing between *backing* and *premises* is not easy because these two components are similar such that they support the *claim*, as discussed in Section 4.4.8. We can also see that the misclassification is consistent among **-B* and **-I* tags.

Rebuttal is often misclassified as *Premise* (28% for *Rebuttal-I* and 18% for *Rebuttal-B*; notice again the consistency in **-B* and **-I* tags). This is rather surprising, as one would expect that *rebuttal* would be confused with a *claim*, because its role is to provide an opposing view.

Refutation-B and *Refutation-I* is misclassified as *Premise-I* in 19% and 27% of the cases, respectively. This finding confirms the discussion in section 4.4.8, because the role of *refutation* is highly context-dependent. From a pragmatic perspective, it is put forward to indirectly support the *claim* by attacking the *rebuttal*, thus having a similar function to the *premise*.

5.1.6 Qualitative Error Analysis. We manually examined misclassified examples produced by the best-performing system to identify which phenomena pose the biggest challenges. Properly detecting boundaries of argument components caused problems, as shown in Figure 9(a). This is in line with the granularity annotation difficulties discussed in Section 4.4.8. The next example in Figure 9(b) shows that even if boundaries of components were detected precisely, the distinction between *premise* and *backing*

Table 12
Probabilistic confusion matrix for the cross-validation scenario for the best performing system from Table 9. Row labels represent gold labels, column labels are predictions. Values are percent; each row sums up to 100%.

	Bac-B	Bac-I	Cla-B	Cla-I	O	Pre-B	Pre-I	Reb-B	Reb-I	Ref-B	Ref-I
Bac-B	28	7	6	0	41	12	5	0	0	0	0
Bac-I	0	35	0	3	36	0	23	0	2	0	1
Cla-B	5	3	24	4	48	10	5	1	1	0	0
Cla-I	0	6	0	26	55	0	13	0	1	0	0
O	0	8	0	2	70	0	18	0	1	0	0
Pre-B	5	4	4	1	49	24	8	2	1	0	0
Pre-I	0	9	0	3	49	0	35	0	4	0	0
Reb-B	3	7	12	4	49	18	3	3	1	0	0
Reb-I	0	9	0	6	44	0	28	0	11	0	1
Ref-B	8	0	0	12	46	8	19	0	4	4	0
Ref-I	0	4	0	4	60	0	27	0	1	0	4

Gold

Some really good points have been expressed here. [...]
[*premise:* We've heard about public school space being allotted to accommodate one religion and its demand for a dedicated space. Muslim prayer is strictly segregated. Gender segregation violates our Charter of Rights and Freedoms which, under Section 15, prohibits discrimination on the grounds of race; national or ethnic origin; colour; religion; gender; age; and mental or physical disability. Sexual orientation has recently been recognized as a prohibited ground for discrimination under the Charter.]
Are gay Muslim students allowed? [...]

Predicted

Some really good points have been expressed here. [...]
[*premise:* We've heard about public school space being allotted to accommodate one religion and its demand for a dedicated space. Muslim prayer is strictly segregated.] [*premise:* Gender segregation violates our Charter of Rights and Freedoms which, under Section 15, prohibits discrimination on the grounds of race; national or ethnic origin; colour; religion; gender; age; and mental or physical disability.] Sexual orientation has recently been recognized as a prohibited ground for discrimination under the Charter.
Are gay Muslim students allowed? [...]

(a) #1346 (article comment, prayer-in-schools)

Gold

Ohhhh, here we go again!!!! Where ever the Muslims go, they expect "special" treatment. [*claim:* No religion in schools....this is what we've come to.] [*premise:* In order to keep everyone satisfied,] [*claim:* there should be no religion in schools.] If parents want their children to have religion, they are going to have to teach them at home or in their places of worship. [...] For goodness sake, it's just awful. [*backing:* Time was, the school day started with a prayer, and yes, maybe religion classes, but not nowadays..there would be full scale war. If the textbooks contain any reference to God, there's trouble, and if the teachers happen to make any kind of a reference to God, or religion..or the hereafter..or what ever may have any religious connotation, the children tell their parents and the parents complain!!!!] So there you have it....yet more proof the multiculturalism doesn't work!! [...]

Predicted

Ohhhh, here we go again!!!! Where ever the Muslims go, they expect "special" treatment. No religion in schools....this is what we've come to. [*claim:* In order to keep everyone satisfied, there should be no religion in schools.] If parents want their children to have religion, they are going to have to teach them at home or in their places of worship. [...] For goodness sake, it's just awful. [*premise:* Time was, the school day started with a prayer, and yes, maybe religion classes, but not nowadays..there would be full scale war. If the textbooks contain any reference to God, there's trouble, and if the teachers happen to make any kind of a reference to God, or religion..or the hereafter..or what ever may have any religious connotation, the children tell their parents and the parents complain!!!!] So there you have it....yet more proof the multiculturalism doesn't work!! [...]

(b) #1412 (artcomment, prayer-in-schools)

Gold

[*claim:* Sending your child to a private school is one of the best things you can do for them.] [*premise:* The teachers do not open up a text book and teach every child the same way. Private teachers have more passion about teaching because they are free to write their own curriculum for each child based on developmental assessments and achievements to where each child is learning at the level they need to be at and not just a class as a whole.] [*premise:* Children in the public school systems ARE left behind, bullied, and not challenged enough in their own learning capabilities.] When your public school system ranks #50 in the nation you tell me, would you rather send your child to public school or honor them with attending a private school?

Predicted

[*claim:* Sending your child to a private school is one of the best things you can do for them.] The teachers do not open up a text book and teach every child the same way. Private teachers have more passion about teaching because they are free to write their own curriculum for each child based on developmental assessments and achievements to where each child is learning at the level they need to be at and not just a class as a whole. [*premise:* Children in the public school systems ARE left behind, bullied, and not challenged enough in their own learning capabilities.] [*claim:* When your public school system ranks #50 in the nation you tell me, would you rather send your child to public school or honor them with attending a private school?]

(c) #2499 (artcomment, public-private-schools)

Gold

[*backing:* I went to both, public and private.] [*premise:* The essential difference were the students. And it makes all the difference.] [*claim:* The public school was a joke.]

Predicted

[*backing:* I went to both, public and private.] The essential difference were the students. And it makes all the difference. The public school was a joke.

(d) #2342 (artcomment, public-private-schools)

Figure 9
Examples of gold data annotations on the left-hand side and system predictions in the best-performing system on the right-hand side.

fails. The example also shows that in some cases, labeling on clause level is required (left-hand side *claim* and *premise*) but the approximation in the system cannot cope with this level of detail (as explained in Section 5.1.1). Confusing non-argumentative text and argument components by the system is sometimes plausible, as is the case of the last rhetorical question in Figure 9(c). On the other hand, the last example in Figure 9(d) shows that some claims using figurative language were difficult to identify. The complete predictions along with the gold data are publicly available.[38]

Hyper-parameter Tuning. SVM^{hmm} offers many hyper-parameters with suggested default values, of which three are of importance. Parameter t sets the order of dependencies of transitions in HMM, parameter e sets the order of dependencies of emissions in HMM, and parameter c represents a trading-off slack versus magnitude of the weight-vector.[39] For all experiments, we set all the hyper-parameters to their default values ($t = 1, e = 0$, $c = 5.0$). Using the best performing feature set from Table 9, we experimented with a grid search over different values ($c \in \{0.1, 1.0, 5.0, 10.0, 50.0\}, e \in \{0, 1\}, t \in \{1, 2, 3\}$) but the results did not outperform the system trained with default parameter values.

5.1.7 Discussion. The F_1 scores might seem very low at first glance. One obvious reason is the actual performance of the system, which gives plenty of room for improvement in the future.[40] But the main cause of low F_1 numbers is the evaluation measure—using 11 classes on the token level is very strict, as it penalizes a mismatch in argument component boundaries the same way as a wrongly predicted argument component type. Therefore we also report two other evaluation metrics that help to put our results into context.

- *Krippendorff's* α_U—This was also used for evaluating inter-annotator agreement (see Section 4.4.6).
- *Boundary similarity* (Fournier 2013)—Using this metric, the problem is treated solely as a segmentation task without recognizing the argument component types.

As shown in Table 13 (the Macro-F_1 scores are repeated from Table 9), the best-performing system achieves a 0.30 score using Krippendorf's α_U, which is in the middle between the baseline and human performance (0.48) but is considered poor from the inter-annotator agreement point of view (Artstein and Poesio 2008). The boundary similarity metrics are not directly suitable for evaluating argument component classification, but reveal a subtask of finding the component boundaries. The best system achieved 0.32 on this measure. Vovk (2013) used this measure to annotate argument spans and his annotators achieved a boundary similarity score of 0.36. Human annotators in Fournier (2013) reached a boundary similarity score of 0.53.

The overall performance of the system is also affected by the accuracy of individual NLP tools used for extracting features. One particular problem is that the preprocessing models we rely on (POS, syntax, semantic roles, coreference, discourse; see Section 5.1.3)

38 https://www.ukp.tu-darmstadt.de/data/argumentation-mining/.

39 http://www.cs.cornell.edu/people/tj/svm_light/svm_hmm.html.

40 We also experimented with a two-step approach consisting of (1) argument component segmentation and (2) argument component classification, but the performance of segmentation (F_1 about 0.7) was not promising.

Table 13
Additional metrics to evaluate the performance of argument component identification applied to the results of 10-fold cross-validation scenario (Table 9). *Measured only on a subset of the data (refer to Section 4.4.6).

	Macro-F_1	**Krippendorff's α_U**	**Boundary similarity**
Human	0.60	0.48*	0.70
Baseline	0.16	0.11	0.18
Best system	0.25	0.30	0.32

were trained on newswire corpora, so one has to expect performance drop when applied to user-generated content. This is, however, a well-known issue in NLP (Foster et al. 2011; Baldwin et al. 2013; Eisenstein 2013).

To obtain an impression of the actual performance of the system on the data, we also provide the complete output of our best performing system in one PDF document together with the gold annotations in the logos dimension side by side in the accompanying software package. We believe this will help the community to see the strengths of our model as well as possible limitations of our current approaches.

6. Conclusions

Let us begin with summarizing answers to the research questions stated in the Introduction. First, as we showed in Section 4.4.2, existing argumentation theories do offer models for capturing argumentation in user-generated content on the Web. We built upon the Toulmin model and proposed some extensions.

Second, as compared to the negative experiences with annotating using Walton's schemes (see Sections 4.4.1 and 3.1), our modified Toulmin model offers a trade-off between its expressiveness and annotation reliability. However, we found that the capabilities of the model to capture argumentation depend on the register and topic, the length of the document, and inherently on the literary devices and structures used for expressing argumentation as these properties influence the agreement among annotators.

Third, there are aspects of online argumentation that lack their established theoretical counterparts, such as rhetorical questions, figurative language, narratives, and fallacies in general. We tried to model some of them in the pathos dimension of argument (Section 4.4.10), but no satisfying agreement was reached. Furthermore, we dealt with a step that precedes argument analysis by filtering documents, given their persuasiveness with respect to the controversy. Finally, we proposed a computational model based on machine learning for identifying argument components (Section 5.1). In this identification task, we experimented with a wide range of linguistically motivated features and found that (1) the largest feature set (including *n*-grams, structural features, syntactic features, topic distribution, sentiment distribution, semantic features, coreference feaures, discourse features, and features based on word embeddings) performs best both in in-domain and all-data cross validation, whereas (2) features based only on word embeddings yield best results in cross-domain evaluation.

Because there is no one-size-fits-all argumentation theory to be applied to actual data on the Web, the argumentation model and an annotation scheme for argumentation mining is a function of the task requirements and the corpus properties. Its selection

should be based on the data at hand and the desired application. Given the proposed use-case scenarios (Section 1) and the results of our annotation study (Section 4.4), we recommend a scheme based on the Toulmin model for short documents, such as comments or forum posts.

Summary of Contributions. In this article we presented our original research of argumentation mining in the user-generated Web discourse by collecting data in six controversial topics in education. We conducted an annotation study on 990 documents to filter persuasive comments and forum posts with inter-annotator agreement of Fleiss' $\kappa = 0.59$. Then we annotated 340 documents (approx. 90k tokens) on the token level with a modified Toulmin model and reached inter-annotator agreement of 0.48 (Krippendorff's joint α_U). We proposed a sequence labeling approach to identify argument components in the discourse and significantly ($p < 0.001$) outperformed the baseline (0.156) with overall Macro-$F_1 = 0.251$. We also found that a feature set based on word embeddings works well in a cross-domain scenario and reaches Macro-$F_1 = 0.209$. We thoroughly examined errors made by the system and proposed future improvements.

As the argumentation mining field is still evolving, and to foster future research, we provide our annotation guidelines, the annotated data, the source codes for the experiments, as well as the results of our system for error analysis. We believe that keeping the whole process transparent will help to identify the strengths and possible shortcomings and will motivate the community to build upon our work.

Appendix A. Raw Corpus Compilation

Given the six controversial topics and four different registers introduced in Section 4.1, we compiled the *raw corpus* semi-automatically. Web sites with relevant comments to articles and discussion forums were identified manually (Google search engine) in order to maximize their relatedness to the search topic. We did not prefer any particular platform or data source. We extracted the texts automatically, however, we did some minimal data pre-selection and cleaning. If article comments formed a tree structure, we kept only the root comments, as they most likely comment on the topic of the article, according to our observations. In discussion forum posts, we automatically removed all quotations (users usually quote the previous post to which they react).

Articles and blogs were also selected manually; we skimmed the texts quickly to check if they discuss the given topic in an argumentative manner. Because we wanted to ensure the reliability of the extracted texts in terms of proper paragraph formatting and boiler-content removal, we extracted the texts manually. For each document we also retained the paragraph formatting, as paragraphs play an important role in argumentative discourse (McGee 2014).

The top ten source domains from the total number of 117 unique domains are listed in Table A.1. Table A.2 shows the document distribution with respect to the registers and topics.

Appendix B. Examples of Annotated Documents from the Second Annotation Study

Example (B.1)
An example of argument annotation with a re-stated claim and both dimensions (logos, pathos). With the first sentence ("Depriving your child of a basic education...") the author appeals to emotions and uses figurative language ("child abuse," "ruin whole

Table A.1
Top 10 source domains in the *raw corpus*.

Domain	Docs	Domain	Docs
living.msn.com	2040	www.theage.com.au	196
discussion.theguardian.com	494	www.forerunner.com	169
community.babycenter.com	403	www.netmums.com	117
www.washingtonpost.com	398	schoolsofthought.blogs.cnn.com	117
www.cbc.ca	380	www.greatschools.org	89

Table A.2
Raw corpus statistics—number of documents for particular topics and registers.

Topic \ Register	**Comment**	**Article**	**Blog**	**Forum post**	**Total**
Redshirting	237	10	10	178	435
Single sex education	237	10	10	76	333
Prayer in schools	547	10	11	240	808
Homeschooling	907	10	11	339	1267
Mainstreaming	33	12	10	134	189
Public vs. private	2235	10	10	157	2412
Total	4196	62	62	1124	5444

life"). The argument is extracted under the original text. Phrases in italics summarize the content of the respective argument components produced by annotators.

Doc#45 (comment, homeschooling) [*app-to-emot:* Depriving your child of a basic education is a form of child abuse. It can ruin your child's whole life.]¶
[*claim:* Home schooling should be illegal] unless [*rebuttal:* the parent can demonstrate that they are providing the same level of education as a public school.] There should be a core national curriculum and testing to ensure children are achieving at least a basic level of education.¶
[*premise:* In an increasingly complex, global technological society, all people need to have a basic understanding of science, technology and local and global culture, just to be able to function and make informed decisions.]¶
[*claim:* I don't see any need for home schooling any child] unless [*rebuttal:* the child has special needs or learning difficulties.] If public schools are under-performing, then the public education system needs to be improved. [*premise:* Public education in the US seems to be a self-perpetuating disaster, with ignorant, uneducated, unqualified people on school boards deciding what children should learn.]

Claim • "Home schooling should be illegal" • "I don't see any need for home schooling any child" **Premise** • *Science and technology are not taught in HS* "In an increasingly complex, global technological society, all people need to have a basic understanding of science, technology and local and global culture, just to be able to function and make informed decisions." • *Public school in the US is bad* "Public education in the US seems to be a self-perpetuating disaster, with ignorant, uneducated, unqualified people on school boards deciding what children should learn." **Rebuttal** • *HS is ok if parents demonstrate the same level of education as in schools* "the parent can demonstrate that they are providing the same level of education as a public school." • *HS can be allowed for kids with special needs* "the child has special needs or learning difficulties." **Appeal to**

emotion • "Depriving your child of a basic education is a form of child abuse. It can ruin your child's whole life."

Example (B.2)
This example contains annotations both in the logos and the pathos dimensions. The main support for the authors' implicit claim starts with "I personally am acquainted with four families ..." Another reason is the social skills.

Doc#163 (comment, homeschooling) Thank you for bringing this tragedy to light. [*backing:* I am a Christian, an educator, a student and a parent and I have seen too many children like the Powells. As an admissions officer, we had applicants whose "record keeping" consisted of sending boxes full of paper for our office to review as part of the application.] [*premise:* If their students did get an interview, which was rare, they didn't have the social skills to survive the first round.]¶
[*premise:* I personlly am acquainted with four families who are home schooling their large families. All four have no intention of book-schooling their daughters past age 13 as they need to learn :homemaking skills". One of the girls, who has not been taught for two years, could be Josh Powell's twin. She is intelligent and desperate to learn, but her parents won't allow it.] [*app-to-emot:* It is heartbreaking.¶
That the Commonwealth of Virginia has such a rich tradition of the education of young people and allows this travesty is shameful.] All of us, no matter our religious beliefs, need to pray that the law changes before more smart children are left behind.

Claim • *Implicit: Against homeschooling* **Premise** • *HS kids lacked social skills* "If their students did get an interview, which was rare, they didn't have the social skills to survive the first round." • *I know families that HS but in fact do not teach their children at all* "I personlly am acquainted with four families who are home schooling their large families. All four have no intention of book-schooling their daughters past age 13 as they need to learn :homemaking skills". One of the girls, who has not been taught for two years, could be Josh Powell's twin. She is intelligent and desperate to learn, but her parents won't allow it." **Backing** • *Observations as an admission officer* "I am a Christian, an educator, a student and a parent and I have seen too many children like the Powells. As an admissions officer, we had applicants whose "record keeping" consisted of sending boxes full of paper for our office to review as part of the application." **Appeal to emotion** • "It is heartbreaking. That the Commonwealth of Virginia has such a rich tradition of the education of young people and allows this travesty is shameful."

Example (B.3)
Notice the wrong capitalization and punctuation. This text had to be annotated on the token level, as the automatic sentence splitting could not cope with it properly.

Doc#2488 (comment, public-private-schools) BIG E.[*premise:* what about all the money we do send to our schools .does it help our child. no .teachers keep asking for more with no difference in teaching just more money and if they dont get it what happens they strike.]well thats real nice on kids education is it not boo hoo you [*claim:* TAKE YOUR KIDS PRIVATE IF YOU CARE AS I DID]

Claim • "TAKE YOUR KIDS PRIVATE IF YOU CARE AS I DID" **Premise** *Teachers in public just want more money but it does not help the kids education* "what about all the money we do send to our schools .does it help our child. no .teachers keep asking for more with no difference in teaching just more money and if they dont get it what happens they strike ."

Example (B.4)
This argument has been annotated as completely in the pathos dimension by only appealing to emotions ("send children to a pig farm").

> **Doc#2581 (comment, public-private-schools)** [*app-to-emot:* Absolutely stupid person, why do we have children? To send to an immoral government run pig farm? No but to give to our children all the best that we as parents can!]
>
> **Claim** • *Implicit: Against public schools* **Appeal to emotion** • "Absolutely stupid person, why do we have children? To send to an immoral government run pig farm? No but to give to our children all the best that we as parents can!"

Acknowledgments

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant I/82806, the German Institute for Educational Research (DIPF), and the German Research Foundation via the German-Israeli Project Cooperation (DIP, grant DA 1600/1-1). Computational resources were provided by the MetaCentrum under the program LM2010005 and the CERIT-SC under the program Centre CERIT Scientific Cloud, part of the Operational Program Research and Development for Innovations, reg. no. CZ.1.05/3.2.00/08.0144. We would like to thank the anonymous reviewers for their valuable feedback and Judith Eckle-Kohler, Christian Stab, Emily Jamison, and Miloslav Konopik for their comments.

References

Aharoni, Ehud, 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, MD.

Ammari, Tawfiq, Meredith Ringel Morris, and Sarita Yardi Schoenebeck. 2014. Accessing social support and overcoming judgment on social media among parents of children with special needs. In *International AAAI Conference on Weblogs and Social Media*, pages 22–31, Ann Arbor, MI.

Amossy, Ruth. 2009. The New Rhetoric's inheritance. Argumentation and discourse analysis. *Argumentation*, 23(3):313–324.

Anthony, Robert and Mijung Kim. 2014. Challenges and remedies for identifying and classifying argumentation schemes. *Argumentation*, 29(1):81–113.

Aristotle and George Kennedy (translator). 1991. *On Rhetoric: A Theory of Civil Discourse*. New York: Oxford University Press.

Artstein, Ron and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34:555–596.

Bal, Bal Krishna and Patrick Saint Dizier. 2010. Towards building annotated resources for analyzing opinions and argumentation in news editorials. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, pages 1152–1158, Valletta.

Baldwin, Timothy, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how different social media sources? In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 356–364, Nagoya.

Ball, W. J. 1994. Using Virgil to analyze public policy arguments: A system based on Toulmin's informal logic. *Social Science Computer Review*, 12(1):26–37.

Bentahar, Jamal, Bernard Moulin, and Micheline Bélanger. 2010. A taxonomy of argumentation models used for knowledge representation. *Artificial Intelligence Review*, 33:211–259.

Bernard, Stéphane, Hugo Mercier, and Fabrice Clément. 2012. The power of well-connected arguments: Early sensitivity to the connective because. *Journal of Experimental Child Psychology*, 111(1):128–135.

Betsch, Cornelia, Frank Renkewitz, Tilmann Betsch, and Corina Ulshöfer. 2010. The

influence of vaccine-critical websites on perceiving vaccination risks. *Journal of Health Psychology*, 15(3):446–455.

Bex, Floris, Trevor Bench-capon, and Bart Verheij. 2012. Persuasive precedents. In *Workshop on Computational Models of Narrative*, pages 171–175, Istanbul.

Bex, Floris J. 2011. *Arguments, Stories and Criminal Evidence*, volume 92 of *Law and Philosophy Library*. Springer Netherlands.

Biber, Douglas and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge University Press.

Biran, Or and Owen Rambow. 2011. Identifying justifications in written dialogs by classifying text as argumentative. *International Journal of Semantic Computing*, 5(4):363–381.

Bishop, Jonathan. 2007. Increasing participation in online communities: A framework for human–computer interaction. *Computers in Human Behavior*, 23(4):1881–1893.

Björnsson, C. H. 1968. *Läsbarhet*. Pedagogiskt Utvecklingsarbete vid Stockholms Skolor. 6. Liber.

Blair, J. Anthony. 2004. Argument and its uses. *Informal Logic*, 24:137–151.

Blair, J. Anthony. 2011. Argumentation as rational persuasion. *Argumentation*, 26(1):71–81.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, March.

Boltužić, Filip and Jan Šnajder. 2014. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58, Baltimore, MA.

Bracewell, David B., Marc Tomlinson, and Hui Wang. 2013. Semi-supervised modeling of social actions in online dialogue. In *2013 IEEE Seventh International Conference on Semantic Computing*, pages 168–175, Irvine, CA.

Brilman, Maarten and Stefan Scherer. 2015. A multimodal predictive model of successful debaters or how I learned to sway votes. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pages 149–158, Brisbane.

Burfoot, Clinton, Steven Bird, and Timothy Baldwin. 2011. Collective classification of congressional floor-debate transcripts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1506–1515, Portland, OR.

Cabrio, Elena, Sara Tonelli, and Serena Villata. 2013. From discourse analysis to argumentation schemes and back: Relations and differences. In João Leite, Tran Cao Son, Paolo Torroni, Leon Torre, and Stefan Woltran, editors, *Proceedings of 14th International Workshop on Computational Logic in Multi-Agent Systems*, volume 8143 of *Lecture Notes in Computer Science*, Springer Berlin-Heidelberg, pages 1–17.

Cabrio, Elena and Serena Villata. 2012. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pages 208–212, Jeju Island.

Cardie, Claire, editor. 2015. *Proceedings of the 2nd Workshop on Argumentation Mining*. Association for Computational Linguistics, Denver, CO.

Chambliss, Marilyn J. 1995. Text cues and strategies successful readers use to construct the gist of lengthy written arguments. *Reading Research Quarterly*, 30(4):778–807.

Choi, Jinho D. 2012. *Optimization of Natural Language Processing Components for Robustness and Scalability*. Ph.D. thesis, University of Colorado Boulder, Computer Science and Cognitive Science.

Cinková, Silvie, Martin Holub, and Vincent Kríž. 2012. Managing uncertainty in semantic tagging. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 840–850, Avignon.

Coleman, Meri and T. L. Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60:283–284.

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

Cortes, Corinna and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Cullen, Rowena and Sarah Morse. 2011. Who's contributing: Do personality traits influence the level and type of participation in online communities. In *44th Hawaii International Conference on*

System Sciences (HICSS), pages 1–11, Kauai, HI

Dagan, Ido, Bill Dolan, Bernardo Magnini, and Dan Roth. 2009. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering*, 15(Special Issue 04):i–xvii.

Damer, T. Edward. 2013. *Attacking Faulty Reasoning: A Practical Guide to Fallacy-Free Arguments*. Cengage Learning, Boston, MA.

de Melo Bezerra, Juliana and Celso Massaki Hirata. 2011. Motivation and its mechanisms in virtual communities. In Adriana S. Vivacqua, Carl Gutwin, and Marcos R. S. Borges, editors, *Collaboration and Technology*, volume 6969 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pages 57–72.

Dippold, Doris. 2007. Using speech frames to research interlanguage pragmatics: Facework strategies in L2 German argument. *Journal of Applied Linguistics*, 4(3):285–308.

Dung, Phan Minh. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and *n*-person games. *Artificial Intelligence*, 77(2):321–357.

Dunn, William N. 2011. *Public Policy Analysis*. Pearson.

Eckart de Castilho, Richard and Iryna Gurevych. 2014. A broad-coverage collection of portable NLP components for building shareable analysis pipelines. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT (OIAF4HLT) at COLING 2014*, pages 1–11, Dublin.

Egg, Markus. 2007. Meaning and use of rhetorical questions. In *Proceedings of the Sixteenth Amsterdam Colloquium*, pages 73–78, Amsterdam.

Eisenstein, Jacob. 2013. What to do about bad language on the Internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369, Atlanta, GA.

Erduran, Sibel, Shirley Simon, and Jonathan Osborne. 2004. TAPping into argumentation: Developments in the application of Toulmin's Argument Pattern for studying science discourse. *Science Education*, 88(6):915–933.

Farkas, Richárd, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The CONLL-2010 Shared Task: Learning to detect hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 1–12, Uppsala.

Faulkner, Adam Robert. 2014. *Automated Classification of Argument Stance in Student Essays: A Linguistically Motivated Approach with an Application for Supporting Argument Summarization*. Ph. D. Dissertation, City University of New York.

Feng, Vanessa Wei and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 987–996, Portland, OR.

Ferschke, Oliver. 2014. *The Quality of Content in Open Online Collaboration Platforms: Approaches to NLP-supported Information Quality Management in Wikipedia*. PhD dissertation, Technische Universität Darmstadt, Darmstadt.

Finocchiaro, Maurice A. 2005. *Arguments about Arguments: Systematic, Critical, and Historical Essays In Logical Theory*. Cambridge University Press, Cambridge.

Fisher, Walter. 1987. *Human Communication As Narration*. University of South Carolina Press.

Flesch, Rudolf. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32:221–233.

Flores-Ferrán, Nydia and Kelly Lovejoy. 2015. An examination of mitigating devices in the argument interactions of L2 Spanish learners. *Journal of Pragmatics*, 76:67–86.

Florou, Eirini, 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, Sofia.

Foster, Jennifer, Ozlem Cetinoglu, Joachim Wagner, Joseph Le Roux, Joakim Nivre, Deirdre Hogan, and Josef van Genabith. 2011. From news to comment: Resources and benchmarks for parsing the language of Web 2.0. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 893–901, Chiang Mai.

Fournier, Chris. 2013. Evaluating text segmentation using boundary edit distance. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1702–1712, Sofia.

Frank, Jane. 1990. You call that a rhetorical question? Forms and functions of rhetorical questions in conversation. *Journal of Pragmatics*, 14(5):723–738.

Freeley, Austin J. and David L. Steinberg. 2008. *Argumentation and Debate*. Cengage Learning, Stamford, CT.

Freeman, James B. 1991. *Dialectics and the Macrostructure of Arguments: A Theory of Argument Structure*, volume 10 of *Trends in Linguistics*. De Gruyter.

Freeman, James B. 2011. *Argument Structure: Representation and Theory*, volume 18 of *Argumentation Library*. Springer Netherlands.

Garcia-Mila, Merce, Sandra Gilabert, Sibel Erduran, and Mark Felton. 2013. The effect of argumentative task goal on the quality of argumentative discourse. *Science Education*, 97(4):497–523.

Garcia-Villalba, Maria Paz and Patrick Saint-Dizier. 2014. Argument extraction in opinion analysis: Identifying the reasons behind consumer evaluations. In Murray E. Jennex, editor, *Knowledge Discovery, Transfer, and Management in the Information Age*. IGI Global, pages 186–211.

Georgila, Kallirroi, Ron Artstein, Angela Nazarian, Michael Rushforth, David Traum, and Katia Sycara. 2011. An annotation scheme for cross-cultural argumentation and persuasion dialogues. In *Proceedings of the SIGDIAL 2011 Conference: the 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 272–278, Portland, OR.

Ghosh, Debanjan, Smaranda Muresan, Nina Wacholder, Mark Aakhus, and Matthew Mitsui. 2014. Analyzing argumentative discourse units in online interactions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 39–48, Baltimore, MA.

Gottipati, Swapna, Minghui Qiu, Yanchuan Sim, Jing Jiang, and Noah A. Smith. 2013. Learning topics and positions from Debatepedia. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1858–1868, Seattle, WA.

Goudas, Theodosis, 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*, Springer International Publishing, pages 287–299.

Govier, Trudy. 2010. *A Practical Study of Argumentation*. Wadsworth Cengage Learning, Belmont, CA.

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

Guadagno, Rosanna E. and Robert B. Cialdini. 2002. Online persuasion: An examination of gender differences in computer-mediated interpersonal influence. *Group Dynamics: Theory, Research, and Practice*, 6(1):38–51.

Guo, Jiang, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Revisiting embedding features for simple semi-supervised learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 110–120, Doha.

Habernal, Ivan 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.

Han, Chung-hye. 2002. Interpreting interrogatives as rhetorical questions. *Lingua*, 112(3):201–229.

Harrell, Maralee. 2011a. Argument diagramming and critical thinking in introductory philosophy. *Higher Education Research & Development*, 30(3):371–385.

Harrell, Maralee. 2011b. Understanding, evaluating, and producing arguments: Training is necessary for reasoning skills. *Behavioral and Brain Sciences*, 34:80–81, 4.

Hasan, Kazi Saidul and Vincent Ng. 2012. Predicting stance in ideological debate with rich linguistic knowledge. In *Proceedings of COLING 2012: Posters*, pages 451–460, Mumbai.

Hasan, Kazi Saidul and Vincent Ng. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1348–1356, Nagoya.

Hitchcock, David. 2003. Toulmin's warrants. In Frans H. Van Eemeren, J. Anthony Blair, Charles A. Willard, and A. Francisca Snoeck Henkemans, editors, *Anyone Who Has a View: Theoretical Contributions to the Study of Argumentation*, volume 8 of

Argumentation Library. Springer Netherlands, pages 69–82.
Hoeken, Hans and Karin M. Fikkers. 2014. Issue-relevant thinking and identification as mechanisms of narrative persuasion. *Poetics*, 44:84–99.
Houy, Constantin, 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-13)*, page 7. Menlo Park, CA.
Huang, Francis L. and Marcia A. Invernizzi. 2013. Birthday effects and preschool attendance. *Early Childhood Research Quarterly*, 28(1):11–23.
Ilie, Cornelia. 1999. Question-response argumentation in talk shows. *Journal of Pragmatics*, 31(8):975–999.
Joachims, Thorsten, Thomas Finley, and Chun-Nam John Yu. 2009. Cutting-plane training of structural SVMs. *Machine Learning*, 77(1):27–59.
Johnson, Ralph H. 2000. *Manifest Rationality: A Pragmatic Theory of Argument*. Routledge, Mahwah, NJ.
Kaltenbök, Gunther, Wiltrud Mihatsch, and Stefan Schneider, editors. 2010. *New Approaches to Hedging*. Number 9 in Studies in Pragmatics. Emerald Group Publishing Limited.
Keerthi, S. Sathiya and S. Sundararajan. 2007. CRF versus SVM-struct for sequence labeling. Technical report, Yahoo! Research.
Ketcham, V. A. 1917. *The Theory and Practice of Argumentation and Debate*. Macmillan, New York.
Kluge, Roland. 2014a. Automatic analysis of arguments about controversial educational topics in Web documents, Masters thesis, Ubiquitious Knowledge Processing Lab, TU Darmstadt.
Kluge, Roland. 2014b. *Searching for Arguments: Automatic analysis of arguments about controversial educational topics in Web documents*. AV Akademikerverlag, Saarbrücken, Germany.
Kreutel, Karen. 2007. "I'm not agree with you." ESL learners' expressions of disagreement. *Teaching English as a Second or Foreign Language*, 11(3):1–35.
Krippendorff, Klaus. 2004. *Content Analysis: An Introduction to Its Methodology*. Sage Publications: Thousand Oaks, CA.
Lafferty, John D., 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*, pages 282–289, San Francisco, CA.
Landis, J. Richard and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.
Le, Elisabeth. 2004. Active participation within written argumentation: Metadiscourse and editorialist's authority. *Journal of Pragmatics*, 36(4):687–714.
Le, Quoc and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196, Beijing.
Lee, Heeyoung, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916.
Lee-Goldman, Russell. 2006. A typology of rhetorical questions. Technical Report 1, University of California, Berkeley.
Leyton Escobar, Mariana, P. A. M. Kommers, and Ardion Beldad. 2014. Using narratives as tools for channeling participation in online communities. *Computers in Human Behavior*, 37:64–72.
Lin, Ziheng, Hwee Tou Ng, and Min-Yen Kan. 2014. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 20(02):151–184.
Lippi, Marco and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology*, 16(2):1–25.
Llewellyn, Clare, Claire Grover, Jon Oberlander, and Ewan Klein. 2014. Re-using an argument corpus to aid in the curation of social media collections. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 462–468, Reykjavik.
Macagno, Fabrizio and Aikaterini Konstantinidou. 2012. What students' arguments can tell us: Using argumentation schemes in science education. *Argumentation*, 27(3):225–243.
MacEwan, E. J. 1898. *The Essentials of Argumentation*. D. C. Heath, Boston, MA.
Madnani, Nitin, Michael Heilman, Joel Tetreault, and Martin Chodorow. 2012. Identifying high-level organizational elements in argumentative discourse. In

Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 20–28, Montreal.

Mann, William C. and Sandra A. Thompson. 1987. Rhetorical structure theory: A theory of text organization. Technical report 1S1/RS-87-185, Information Sciences Institute, University of Southern California, Marina del Rey, CA.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, MA.

McCallum, Andrew Kachites. 2002. MALLET: A Machine Learning for Language Toolkit. mallet.cs.umass.edu.

McGee, Iain. 2014. The pragmatics of paragraphing English argumentative text. *Journal of Pragmatics*, 68:40–72.

Mercier, Hugo and Dan Sperber. 2011. Why do humans reason? Arguments for an argumentative theory. *The Behavioral and Brain Sciences*, 34(2):57–74; discussion 74–111.

Miceli, Maria, Fiorella de Rosis, and Isabella Poggi. 2006. Emotional and non-emotional persuasion. *Applied Artificial Intelligence*, 20(10):849–879.

Micheli, Raphaël. 2008. Emotions as objects of argumentative constructions. *Argumentation*, 24(1):1–17.

Micheli, Raphaël. 2011. Arguing without trying to persuade? Elements for a non-persuasive definition of argumentation. *Argumentation*, 26(1):115–126.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc., pages 3111–3119.

Mochales, Raquel and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

Mohammadi, Gelareh, Sunghyun Park, Kenji Sagae, Alessandro Vinciarelli, and Louis-Philippe Morency. 2013. Who is persuasive? The role of perceived personality and communication modality in social multimedia. In *Proceedings of the 15th ACM International Conference on Multimodal Interaction*, ICMI '13, pages 19–26, Sydney.

Murphy, P. Karen. 2001. What makes a text persuasive? Comparing students' and experts' conceptions of persuasiveness. *International Journal of Educational Research*, 35(7-8):675–698.

Nettel, Ana Laura and Georges Roque. 2011. Persuasive argumentation versus manipulation. *Argumentation*, 26(1):55–69.

Newman, S. and C. Marshall. 1991. Pushing Toulmin too far: Learning from an argument representation scheme. Technical report SSL-92-45, Xerox Palo Alto Research Center, Palo Alto, CA.

Nguyen, Vinh Van, Minh Le Nguyen, and Akira Shimazu. 2009. Clause Splitting with Conditional Random Fields. *Information and Media Technologies*, 4(12):57–75.

Nicholson, Michelle S. and Julie Leask. 2012. Lessons from an online debate about measles-mumps-rubella (MMR) immunization. *Vaccine*, 30(25):3806–3812.

Niehaus, James, Victoria Romero, Jonathan Pfautz, Scott Neal Reilly, Richard Gerrig, and Peter Weyhrauch. 2012. Towards a computational model of narrative persuasion: A broad perspective. In *Workshop on Computational Models of Narrative*, pages 181–182, Istanbul.

Nieminen, Petteri and Anne-Mari Mustonen. 2014. Argumentation and fallacies in creationist writings against evolutionary theory. *Evolution: Education and Outreach*, 7(1):11.

O'Keefe, Daniel J. 1982. The concepts of argument and arguing. In J. R. Cox and C. A. Willard, *Advances in Argumentation Theory and Research*, Southern Illinois University Press, pages 3–23.

O'Keefe, Daniel J. 2002. *Persuasion: Theory and Research*. Sage Publications.

O'Keefe, Daniel J. 2011. Conviction, persuasion, and argumentation: Untangling the ends and means of influence. *Argumentation*, 26(1):19–32.

Ong, Nathan, Diane Litman, and Alexandra Brusilovsky. 2014. Ontology-based argument mining and automatic essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 24–28, Baltimore.

Ottati, Victor, Susan Rhoads, and Arthur C. Graesser. 1999. The effect of metaphor on processing style in a persuasion task: A motivational resonance model. *Journal of Personality and Social Psychology*, 77(4):688.

Paglieri, Fabio and Cristiano Castelfranchi. 2014. Trust, relevance, and arguments. *Argument & Computation*, 5(2-3):216–236.
Park, Joonsuk 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, MD.
Park, Souneil, Kyung Soon Lee, and Junehwa Song. 2011. Contrasting opposing views of news articles on contentious issues. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 340–349, Portland, OR.
Peldszus, Andreas. 2014. Towards segment-based recognition of argumentation structure in short texts. In *Proceedings of the First Workshop on Argumentation Mining*, pages 88–97, Baltimore, MD.
Peldszus, Andreas and Manfred Stede. 2013a. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence*, 7(1):1–31.
Peldszus, Andreas and Manfred Stede. 2013b. Ranking the annotators: An agreement study on argumentation structure. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 196–204, Sofia.
Peldszus, Andreas and Manfred Stede. 2015. Joint prediction in MST-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948, Lisbon.
Perkins, David N. 1985. Postprimary education has little impact on informal reasoning. *Journal of Educational Psychology*, 77(5):562–571.
Petty, Richard E., John T. Cacioppo, and Martin Heesacker. 1981. Effects of rhetorical questions on persuasion: A cognitive response analysis. *Journal of Personality and Social Psychology*, 40(3):432–440.
Pineau, Andrew. 2013. The abuses of argument: Understanding fallacies on Toulmin's layout of argument. In D. Mohammed and M. Lewiński, editors, *Virtues of Argumentation. Proceedings of the 10th International Conference of the Ontario Society for the Study of Argumentation (OSSA)*, volume 33. OSSA, Winsdor, CA, pages 1–11.
Platt, John C. 1999. Fast training of support vector machines using sequential minimal optimization. In Bernhard Schölkopf, Christopher J. C. Burges, and Alexander J. Smola, editors, *Advances in Kernel Methods: Support Vector Learning*. MIT Press, Cambridge, MA, pages 185–208.
Prakken, Henry and Gerard Vreeswijk. 2002. Logics for defeasible argumentation. In D. M. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, volume 4 of *Handbook of Philosophical Logic*. Springer Netherlands, pages 219–318.
Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 1–4, Marrakech.
Procter, Rob, Farida Vis, and Alex Voss. 2013. Reading the riots on Twitter: Methodological innovation for the analysis of big data. *International Journal of Social Research Methodology*, 16(3):197–214.
Qiu, Minghui and Jing Jiang. 2013. A latent variable model for viewpoint discovery from threaded forum posts. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1031–1040, Atlanta, GA.
Qiu, Minghui, Liu Yang, and Jing Jiang. 2013. Modeling interaction features for debate side clustering. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, pages 873–878, Burlingame, CA.
Rapanta, Chrysi, Merce Garcia-Mila, and Sandra Gilabert. 2013. What is meant by argumentative competence? An integrative review of methods of analysis and assessment in education. *Review of Educational Research*, 83(4):483–520.
Rapp, Christof and Tim Wagner. 2012. On some Aristotelian sources of modern argumentation theory. *Argumentation*, 27(1):7–30.
Reed, Chris and Glenn Rowe. 2004. Araucaria: Software for argument analysis, diagramming and representation. *International Journal on Artificial Intelligence Tools*, 13(04):961–979.
Reed, Chris and Glenn Rowe. 2006. Translating Toulmin diagrams: Theory neutrality in argument representation. *Argumentation*, 19(3):267–286.
Reed, Chris and Douglas Walton. 2003. Argumentation schemes in argument-as-process and argument-as-product. In *Proceedings*

of the Conference Celebrating Informal Logic, pages 1–11, Windsor, Ontario.
Roberts, Richard M. and Roger J. Kreuz. 1994. Why do people use figurative language? *Psychological Science*, 5(3):159–163.
Rooney, Niall, Hui Wang, and Fiona Browne. 2012. Applying kernel methods to argumentation mining. In *Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference*, pages 272–275, Marco Island, FL.
Rosenthal, Sara and Kathleen McKeown. 2012. Detecting opinionated claims in online discussions. In *2012 IEEE Sixth International Conference on Semantic Computing*, pages 30–37, Palermo.
Saint-Dizier, Patrick. 2012. Processing natural language arguments with the TextCoop platform. *Argument & Computation*, 3(1):49–82.
Santibáñez, Cristián. 2010. Metaphors and argumentation: The case of Chilean parliamentarian media participation. *Journal of Pragmatics*, 42:973–989.
Scheuer, Oliver, Frank Loll, Niels Pinkwart, and Bruce M. McLaren. 2010. Computer-supported argumentation: A review of the state of the art. *International Journal of Computer-Supported Collaborative Learning*, 5(1):43–102.
Schiappa, Edward and John P. Nordin. 2013. *Argumentation: Keeping Faith with Reason*. Pearson UK.
Schlosser, Ann E. 2011. Can including pros and cons increase the helpfulness and persuasiveness of online reviews? The interactive effects of ratings and arguments. *Journal of Consumer Psychology*, 21(3):226–239.
Schmidt-Radefeldt, Jürgen. 1977. On so-called 'rhetorical' questions. *Journal of Pragmatics*, 1(4):375–392.
Schneider, Jodi, Brian Davis, and Adam Wyner. 2012. Dimensions of argumentation in social media. In Annette ten Teije, Johanna Völker, Siegfried Handschuh, Heiner Stuckenschmidt, Mathieu d'Acquin, Andriy Nikolov, Nathalie Aussenac-Gilles, and Nathalie Hernandez, editors, *Knowledge Engineering and Knowledge Management*, volume 7603 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pages 21–25.
Schneider, Jodi, Tudor Groza, and Alexandre Passant. 2013. A review of argumentation for the Social Semantic Web. *Semantic Web*, 4(2):159–218.
Schneider, Jodi, Krystian Samp, Stefan Decker, and Alexandre Passant. 2013. Arguments about deletion: How experience improves the acceptability of arguments in ad-hoc online task groups. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, pages 1069–1079, San Antonio, TX.
Schneider, Jodi and Adam Wyner. 2012. Identifying consumers' arguments in text. In Diana Maynard, Marieke van Erp, and Brian Davis, editors, *Semantic Web and Information Extraction SWAIE 2012*, pages 31–42, Galway City.
Senter, J. R. and E. A. Smith. 1967. Automated readability index. Technical report AMRL-TR-66-220, Aerospace Medical Research Laboratories, Ohio.
Sergeant, Alan. 2013. Automatic argumentation extraction. In *ESWC 2013*, pages 656–660, Montpellier.
Shaw, Victoria F. 1996. The cognitive processes in informal reasoning. *Thinking & Reasoning*, 2(1):51–80.
Simosi, Maria. 2003. Using Toulmin's framework for the analysis of everyday argumentation: Some methodological considerations. *Argumentation*, 17:185–202.
Smirnova, Alla Vitaljevna. 2009. Reported speech as an element of argumentative newspaper discourse. *Discourse & Communication*, 3:79–103.
Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, WA.
Somasundaran, Swapna and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 226–234, Suntec.
Stab, Christian and Iryna Gurevych. 2014a. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510, Dublin.
Stab, Christian 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), pages 46–56, Doha.
Stegmann, Karsten, Christof Wecker, Armin Weinberger, and Frank Fischer. 2011. Collaborative argumentation and cognitive elaboration in a computer-supported collaborative learning environment. *Instructional Science*, 40(2):297–323.
Sun, Na, Patrick Pei-Luen Rau, and Liang Ma. 2014. Understanding lurkers in online communities: A literature review. *Computers in Human Behavior*, 38(0):110–117.
Teninbaum, Gabriel H. 2011. Who cares? *Drexel Law Review*, 3:485–519.
Teufel, Simone, Jean Carletta, and Marc Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, pages 110–117, Bergen.
Teufel, Simone and Marc Moens. 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28:409–445.
Teufel, Simone, Advaith Siddharthan, and Colin Batchelor. 2009. Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1493–1502.
Thomas, Stephen N. 1981. *Practical Reasoning in Natural Language*. Prentice-Hall, Englewood Cliffs, NJ.
Tindale, Christopher W. 2007. *Fallacies and Argument Appraisal*. Cambridge University Press, New York, NY.
Tjong, Erik F., Kim Sang, and Hervé Déjean. 2001. Introduction to the CoNLL-2001 shared task: Clause identification. In *Proceedings of the ACL 2001 Workshop on Computational Natural Language Learning (ConLL)*, pages 5–9, Toulouse.
Toulmin, Stephen, Richard Rieke, and Allan Janik. 1984. *An Introduction to Reasoning*. Macmillan, New York.
Toulmin, Stephen E. 1958. *The Uses of Argument*. Cambridge University Press.
Toulmin, Stephen E. 2003. *The Uses of Argument, Updated Edition*. Cambridge University Press, New York.
Trabelsi, Amine and Osmar R. Zaiane. 2014. Finding arguing expressions of divergent viewpoints in online debates. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, pages 35–43, Gothenburg.
van Eemeren, Frans H., Bart Garssen, Erik C. W. Krabbe, A. Francisca Snoeck Henkemans, Bart Verheij, and Jean H. M. Wagemans. 2014. *Handbook of Argumentation Theory*. Springer, Berlin/Heidelberg.
van Eemeren, Frans H., R. Grootendorst, and T. Kruiger. 1987. *Handbook of Argumentation Theory: A Critical Survey of Classical Backgrounds and Modern Studies*. Foris Publications.
van Eemeren, Frans H., R. Grootendorst, and A. F. Snoeck Henkemans. 2002. *Argumentation: Analysis, Evaluation, Presentation*. Lawrence Erlbaum, Mahwah, NJ.
van Eemeren, Frans H. and Rob Grootendorst. 1984. *Speech Acts in Argumentative Discussions: A Theoretical Model for the Analysis of Discussions Directed Towards Solving Conflicts of Opinion*, volume 1. Foris Publications.
Villalba, Maria Paz Garcia and Patrick Saint-Dizier. 2012. Some facets of argument mining for opinion analysis. In *Proceedings of Fourth International Conference on Computational Models of Argument, COMMA 2012*, pages 23–34, Vienna.
Voss, James F. 2006. Toulmin's model and the solving of ill-structured problems. *Argumentation*, 19(3):321–329.
Vovk, Artem. 2013. Discovery and analysis of public opinions on controversial topics in the educational domain. Master's Thesis, Ubiquitious Knowledge Processing Lab, TU Darmstadt.
Wacholder, Nina, Smaranda Muresan, Debanjan Ghosh, and Mark Aakhus. 2014. Annotating multiparty discourse: Challenges for agreement metrics. In *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*, pages 120–128, Dublin.
Wachsmuth, Henning, Martin Trenkmann, Benno Stein, Gregor Engels, and Tsvetomira Palakarska. 2014. A review corpus for argumentation analysis. In *15th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 14)*, pages 115–127, Kathmandu.
Walker, Marilyn, Pranav Anand, Rob Abbott, and Ricky Grant. 2012. Stance classification using dialogic properties of persuasion. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 592–596, Montréal.

Walton, Douglas. 2005. *Fundamentals of Critical Argumentation*. Critical Reasoning and Argumentation. Cambridge University Press.

Walton, Douglas. 2007. *Dialog Theory for Critical Argumentation*. John Benjamins Publishing Company.

Walton, Douglas. 2012. Using argumentation schemes for argument extraction: A bottom–up method. *International Journal of Cognitive Informatics and Natural Intelligence*, 6(3):33–61.

Walton, Douglas, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press.

Wang, Xiao-Ning, Jin-Mao Wei, Han Jin, Gang Yu, and Hai-Wei Zhang. 2013. Probabilistic confusion entropy for evaluating classifiers. *Entropy*, 15(11):4969–4992.

Webber, Bonnie, Markus Egg, and Valia Kordoni. 2012. Discourse structure and language technology. *Natural Language Engineering*, 18(04):437–490.

Weinberger, Armin and Frank Fischer. 2006. A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education*, 46(1):71–95.

Weinstock, Michael, Yair Neuman, and Iris Tabak. 2004. Missing the point or missing the norms? Epistemological norms as predictors of students' ability to identify fallacious arguments. *Contemporary Educational Psychology*, 29(1):77–94.

Wolfe, Christopher R., M. Anne Britt, and Jodie A. Butler. 2009. Argumentation schema and the myside bias in written argumentation. *Written Communication*, 26(2):183–209.

Xu, Cihua and Yicheng Wu. 2014. Metaphors in the perspective of argumentation. *Journal of Pragmatics*, 62:68–76.

Zhang, Qi, Jin Qian, Huan Chen, Jihua Kang, and Xuanjing Huang. 2013. Discourse level explanatory relation extraction from product reviews using first-order logic. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 946–957, Seattle, WA.