Topic Ontologies for Arguments

Yamen Ajjour
Lebiniz University Hannover

Johannes Kiesel
Bauhaus-Universität Weimar

Benno Stein
Bauhaus-Universität Weimar

Martin Potthast
Leipzig University and ScaDS.AI

Abstract

Many computational argumentation tasks, such as stance classification, are topic-dependent: The effectiveness of approaches to these tasks depends largely on whether they are trained with arguments on the same topics as those on which they are tested. The key question is: What are these training topics? To answer this question, we take the first step of mapping the argumentation landscape with The Argument Ontology (TAO). TAO draws on three authoritative sources for argument topics: the World Economic Forum, Wikipedia's list of controversial topics, and Debatepedia. By comparing the topics in our ontology with those in 59 argument corpora, we perform the first comprehensive assessment of their topic coverage. While TAO already covers most of the corpus topics, the corpus topics barely cover all the topics in TAO. This points to a new goal for corpus construction to achieve a broad topic coverage and thus better generalizability of computational argumentation approaches.

1 Introduction

The term "topic" refers to the subject matter of a text. A text may be about one or more topics and the relationship between topics and texts is called "aboutness" (Yablo, 2014). Topics play a central role in argumentation because they determine argumentation strategies and rhetorical devices by setting the appropriate and expected universe of discourse. This view is supported by pragma-dialectics (van Eemeren, 2015): "The basic aspects of strategic maneuvering [. . .] are making an expedient selection from the 'topical potential' available at a certain discussion." Although debaters often use commonplace arguments across topics (Bilu et al., 2019), they must be relevant: a black market argument, for example, can be equally well applied to topics such as banning drugs or banning firearms. As recently shown, for example, by Reuver et al. (2021), training computational models to extract, analyze, or generate arguments with a broad topic coverage improves their generalizability.

A set of topics can be organized as a graph, sometimes called a "topic space". Information theorists and library scientists map hierarchical subject relationships into ontologies in this way (Hjørland, 2001). For this purpose, topics are labeled with a subject heading, a phrase from a controlled vocabulary that describes a topic in a concise and discriminating manner. While library ontologies are not focused on argumentation, others deal specifically with *argumentative* topic spaces. We have identified and tapped three authoritative sources of ontological knowledge covering global issues, controversies, and popular debates: the World Economic Forum's "Strategic Intelligence" site, Wikipedia's list of controversial topics, and Debatepedia's debate classification system (Section 4). They form the basis for The Argument Ontology (TAO).[1]

We compile a comprehensive survey of 59 argument corpora (Section 3) and investigate their topic coverage with respect to the three authoritative ontologies (Section 5). The coverage of corpora with topic labels is manually assessed by matching each label with the topics of the ontologies. From this, the ontology topics covered by a corpus and the distribution of corpus arguments in the ontologies are calculated. Our analyses show that the existing corpora focus on only a subset of the known topics. For corpora without topic labels, we categorize their argumentative texts by measuring their semantic relatedness to ontology topics. Given the large number of ontology topics (748 for Wikipedia), this is a challenging classification for which we achieve a remarkable F_1 of 0.59. (Section 6).[2]

Altogether, we lay the foundation for the study and systematic exploration of controversial topics within computational argumentation analysis. The authoritative sources identified already cover their respective areas quite comprehensively. Future work will need to extend our approach to other subject areas, such as business, domestic, historical, and scientific argument spaces.

[1] Data: https://zenodo.org/record/3928096.

[2] Code: https://github.com/webis-de/EACL-23.

2 Related Work

Our review of related work focuses on the role of the variable "topic" in computational argumentation. Moreover, we briefly review topic ontologies and hierarchical topic classification.

2.1 Topics in Computational Argumentation

In computational argumentation, arguments are typically modeled as compositions of argument units, where an argument unit is represented as a span of text. Habernal and Gurevych (2016a) adopts Toulmin (1958)'s (1958) model, which defines six unit types, among which are "claim" and "data". Wachsmuth et al. (2017) employ a more basic model of two units, which defines an argument as a claim or conclusion supported by one or more premises. These models capture arguments without explicitly identifying the topic they address. Levy et al. (2014) consider claims to be topic-dependent and study their detection in the context of a random selection of 32 topics from idebate.org. This work raises the question why topic-dependence has not been addressed more urgently until now.

Key tasks for computational argumentation include the mining of arguments from natural language (Moens et al., 2007; Al-Khatib et al., 2016), classifying their stances with regard to a thesis (Bar-Haim et al., 2017), and analyzing which arguments are more persuasive (Tan et al., 2016; Habernal and Gurevych, 2016a). Current approaches to these tasks rely on supervised classification. Daxenberger et al. (2017) show that supervised classifiers fail to generalize across domains ($\sim$ topics). More recently, Stab et al. (2018) tweak BiLSTM (Graves and Schmidhuberab, 2005) to integrate the topic while jointly detecting (1) whether a sentence is an argument and (2) its stance to the topic. The designed neural network outperforms BiLSTM without topic integration in both tasks; further evidence for the topic-dependence of argument mining and stance classification. Whether model transfer between more closely related topics works better is unknown. As a first step, Reuver et al. (2021) show that cross-topic stance-classification with BERT (Devlin et al., 2018) produces mixed results depending on the topics, but misses the relations between the topics. Gu et al. (2018) show that integrating the topic of an argument helps assessing its persuasiveness.

Topic plays a central role in argument retrieval and generation since it defines what arguments are relevant. Argument retrieval aims at delivering pro and con arguments on a given topic query. A major challenge in argument retrieval is the grouping of arguments that address common aspects of a topic. As shown by Reimers et al. (2019) and Ajjour et al. (2019a), integrating the topic is an important step while clustering arguments. For argument generation, Bilu et al. (2019) introduce an approach that matches an input topic against a list of topics that are paired with sets of topic-adjustable commonplace arguments (e.g., black-market arguments). In a similar vein, Bar-Haim et al. (2019) identify consistent and contrastive topics for a given topic with the goal of expanding the topic in a new direction (e.g., fast food versus obesity). Both approaches show the merit of utilizing argument topic ontologies in argument generation.

Only abstract argumentation may be truly topic-independent, where only the structure and relations among arguments, not their language, are studied.

2.2 Topic Ontologies

In information science, an ontology is defined as "an explicit specification of a conceptualization" (Gruber, 1993). Topic ontologies are a specific type of ontologies which specify topics as nodes of a directed acyclic graph. An edge in the graph then implies an "is part of"-relation between the topics (Xamena et al., 2017). The effort in creating topic ontologies ranges from ad-hoc decisions (e.g., tags for blog posts) to extensive classification schemes for libraries. The oldest classification scheme that is still used today in libraries is the Dewey Decimal Classification. It has been translated into over 30 languages, and it contains several tens of thousands of classes. Most topic ontologies focus on a specific domain, such as a the ACM Computing Classification System for computer science, or DMOZ for web pages.[3] The only topic ontology directly linked to arguments is that of Debatepedia.

2.3 Hierarchical Text Classification

Hierarchical text classification aims at classifying a document into a class hierarchy. Depending on how the hierarchical structure is exploited, classification can be done top-down (from higher classes downwards), bottom-up, or flat (ignoring hierarchical relations) (Silla and Freitas, 2011). Researchers usually train supervised classifiers for each class in the hierarchy (Sun and Lim, 2001).

[3] https://dl.acm.org/ccs and https://dmoz-odp.org/

Corpus	Authors	Source	Unit granularity	Units	Topics	Exp.
		Manual selection				
Arguing Subjectivity	Conard et al. (2012)	Editorials	Editorial/blog	84	1	1
Arguments Moderation	Falk et al. (2021)	Discussion forum	Argument	112	2	2
Argumentative Sentences	Eyal et al. (2020)	Wikipedia	Arguments	700	20	1
Argument Facet Similarity	Misra et al. (2016)	Debate portals	Argument	6,188	3	12
AURC	Trautmann et al. (2020)	Web	Argument Unit	8,000	8	6
Basn	Kondo et al. (2021)	Debate portals	Argument pair	2,370	6	1
CCSA	Li et al. (2022)	Scientific papers	Argument unit	18,332	1	1
Claim and Evidence 1	Aharoni et al. (2014)	Wikipedia	Wikipedia article	315	33	22
Claim and Evidence 2	Rinott et al. (2015)	Wikipedia	Wikipedia article	547	58	16
Claim Generation	Gretz et al. (2020)	Generated text	Argument Unit	2,839	136	1
Claim Stance	Bar-Haim et al. (2017)	Wikipedia	Argument Unit	2,394	55	15
Claim Sentence Search	Levy et al. (2018)	Wikipedia	Argument unit	1,492,077	150	5
COMARG	Boltužić and Šnajder (2014)	Debate portals	Argument pair	2,298	2	3
Evidence Sentences	Schnarch et al. (2018)	Wikipedia	Argument unit	5,783	118	6
Evidence Sentences 2	Ein-Dor et al. (2020)	Wikipedia	Argument unit	29,429	221	4
Evidence Quality	Gleize et al. (2019)	Wikipedia	Argument pair	5,697	69	2
IAM	Cheng et al. (2022)	Wikipedia	Argument unit	69,666	100	1
ICLE Essay Scoring	Persing et al. (2010)	Essays	Essay	1,000	10	12
Ideological Debates Reasons	Hasan and Ng (2014)	Debate portals	Argument	4,903	4	12
Internet Argument Corpus v2	Abbott et al. (2016)	Web	Discussion	16,555	19	22
Key Point Analysis	Bar-Haim et al. (2020)	Wikipedia	Argument	24,093	28	15
M-Arg	Mestre et al. (2021)	Presidential debate	Argument pair	4,104	18	1
Micro Text v1	Peldszus and Stede (2015)	Essays	Essay	112	18	13
Micro Text v2	Skeppstedt et al. (2018)	Essays	Essay	171	35	2
Multilingual Argument Mining	Toledo-Ronen et al. (2020)	Wikipedia	Argument unit	65,708	347	4
Political Argumentation	Menini et al. (2018)	Presidential debate	Argument pair	1,462	5	3
Record Debating Dataset 2	Mirkin et al. (2018)	Debating	Speech	200	50	5
Record Debating Dataset 3	Lavee et al. (2019)	Debating	Speech	400	199	1
Record Debating Dataset 4	Orbach et al. (2019)	Debating	Speech	200	50	1
Record Debating Dataset 5	Orbach et al. (2020)	Debating	Speech	3,562	397	1
Sci-arg	Lauscher et al. (2018)	Scientific papers	Paper	40	1	7
SciARK	Fergadis et al. (2021)	Scientific papers	Abstract	1,000	6	1
UKP Sentential	Stab et al. (2018)	Web	Argument	25,492	8	20
UKP Aspect	Reimers et al. (2019)	Web	Argument pair	3,595	28	11
UKPConvArg1	Habernal and Gurevych (2016c)	Debate portals	Argument pair	11,650	16	16
UKPConvArg2	Habernal and Gurevych (2016b)	Debate portals	Argument pair	9,111	16	6
WebDiscourse	Habernal and Gurevych (2016a)	Web	Document	340	6	12
Webis-debate-16	Al-Khatib et al. (2016a)	Debate portals	Debate	445	14	5
VivesDebate	Ruiz-Dolz et al. (2021)	Debating	Debate	29	1	2
		Source-driven: greedy within a time-span				
AIFdb	Bex et al. (2013)	Web	Argument unit	67,408	n/a	8
Args-me	Ajjour et al. (2019b)	Debate portals	Argument	387,692	n/a	31
ChangeMyView	Tan et al. (2016)	Discussion forum	Post/comment	14,066	n/a	37
CJEU	Grundler et al. (2022)	Law Case	Court Decision	40	n/a	1
DebateSum	Roush and Balaji (2020)	Debating	Debate	187,386	n/a	1
IMHO	Chakrabarty et al. (2019)	Discussion forum	Argument Unit	5,569,962	n/a	3
Intelligence Squared Debates	Zhang et al. (2016)	Debate portals	Debate	108	n/a	9
Kialo	Kialo (2020)	Debate portals	Argument unit	331,684	n/a	23
Political Speech	Lippi and Torroni (2016)	Ministerial debate	Argument unit	152	n/a	2
USElecDeb60To16	Haddadan et al. (2019)	Presidential debate	Debate	42	n/a	5
MultiOpEd	Liu et al. (2021)	Editorials	Editorial	2,794	n/a	2
QT30	Hautli-Janisz et al. (2022)	Debating	Argument unit	19,842	n/a	1
Review-Rebuttal	Cheng et al. (2020)	Scientific reviews	Argument pair	4,764	n/a	5

Table 1 (continued on next page).

Table 1 (continued).

Corpus	Authors	Source	Unit granularity	Units	Topics	Exp.
Source-driven: sampled						
Argument Annotated Essays	Stab and Gurevych (2017)	Essays	Essay	402	n/a	64
E-rulemaking	Park and Cardie (2018)	Discussion forum	Argument	731	n/a	9
ECHR	Poudyal et al. (2020)	Law Case	Argument	743	n/a	8
Editorials	Al-Khatib et al. (2016b)	Editorials	Editorial	300	n/a	15
GAQCorpus	Ng et al. (2020)	Web	Argument	6,424	n/a	4
IDebate Persuasiveness	Persing and Ng (2017)	Debate portals	Argument	1,205	n/a	1
Scinf-biomed	Gao et al. (2022)	Scientific papers	Paper	27,924	n/a	1

Table 1: Survey of argument corpora indicating data source, unit granularity, and size in terms of units and topics (if authors remarked on it). The unit granularity is the one in the corpus' files, using premises and conclusions as one unit each and the best context-preserving unit for corpora featuring multiple granularities. We presume these topic selection directives from the corpus description: either *manual selection* by the authors, or *source-driven*—i.e., the topics in the selected source(s)—from the units of a specific *time-span* or by random *sampling*. Experiments (Exp.) denotes the count of papers that use the corpus in an experiment among those papers that cite the corpus' paper.

3 Survey of Argument Corpora

To study arguments and computational argumentation tasks, researchers compile corpora with argumentative texts. To the best of our knowledge, Table 1 compiles all corpora dedicated to argumentation until 2022. We review these corpora and their associated publications with regard to what are the sources of arguments, what is the granularity of the corpus, what is the size of the corpora in terms of their units, and which and how many different topics are covered in them. Reviewing all papers citing a corpus, we also analyzed how many experiments were carried out using them.

The most elaborate discussion of topic selection is given in Habernal and Gurevych (2016a), who chose six topics (homeschooling, public versus private schools, redshirting, prayers in schools, single sex education, mainstreaming) to focus on different education-related aspects. The broadest selection of topics is reported by the researchers of IBM Debater,[4] who obtain arguments from Wikipedia. However, samples of the topics have been used in their papers without mentioning which ones. The only other work mentioning their source of topics stems from Stab et al. (2018), who randomly select 8 topics from two lists of controversial topics that originate from an online library and the debate portal ProCon.org, respectively. Peldszus and Stede (2015) predefine a set of topics and give writers the freedom to choose which one to write about, but nothing is said about where the set of predefined topics originate from. Conard et al. (2012) and Hasan and Ng (2014) explicitly select one and four topics, respectively. For all other corpora with topic labels, their authors do not argue on choosing topics, nor selection or sampling criteria. Neither do the authors of corpora without topic labels.

Altogether, it appears that the best practices in argumentation do not as of yet consider topic sampling as a prerequisite task to ensure coverage of a certain domain of interest, diversity, or reproducibility. Based on our review, we presume three basic topic selection directives are in use today: (1) *Manual selection.* Topics are manually defined or selected. Although the process may be random, when aiming for controversial topics, one may often end up with commonplace topics in Western culture (e.g., abortion, death penalty, gay marriage). Still, they are relevant and important today. (2) *Source-driven (greedy within a time-span).* A source of argument ground truth is either exploited in its entirety, or a maximum subset fulfilling desired properties is used. Since argument-related ground truth is hard to come by, it is understandable that many readily available sources are being exploited. (3) *Source-driven (sampled).* A source or argument ground truth is exploited and a subset is sampled. Here, it may be infeasible to exploit a source in its entirety. Al-Khatib et al. (2016b) randomly select 300 documents from three websites. Park and Cardie (2018) and Stab and Gurevych (2017) do not mention anything about their sampling process. In general, both source-driven corpus construction approaches inevitably incur the source's idiosyncracies of topic selection in terms of skew towards certain topics. Scaling up may or may not be a remedy for this problem.

[4] https://www.research.ibm.com/haifa/dept/vst/debating_data.shtml

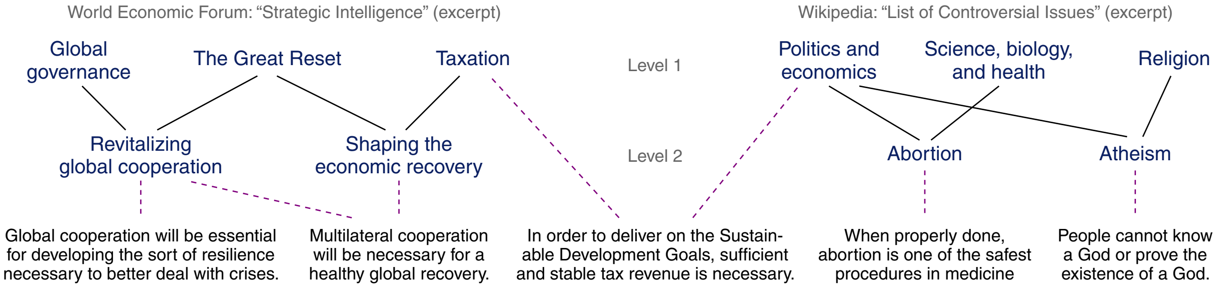

World Economic Forum: "Strategic Intelligence" (excerpt)
Wikipedia: "List of Controversial Issues" (excerpt)
Global governance
The Great Reset
Taxation
Level 1
Politics and economics
Science, biology, and health
Religion
Revitalizing global cooperation
Shaping the economic recovery
Level 2
Abortion
Atheism
Global cooperation will be essential for developing the sort of resilience necessary to better deal with crises.
Multilateral cooperation will be necessary for a healthy global recovery.
In order to deliver on the Sustainable Development Goals, sufficient and stable tax revenue is necessary.
When properly done, abortion is one of the safest procedures in medicine
People cannot know a God or prove the existence of a God.

Figure 1: Example for an assignment of arguments (bottom) to topics of a two-leveled ontology. Level 2 topics are subtopics of their linked Level 1 topics. Arguments linked to a Level 2 topic also pertain to its Level 1 ancestors.

We assess how many experiments have been reported on each of the corpora by collecting the publications referring to a corpus as per Google Scholar, focusing on conference and journal papers, but excluding books and web pages. We then check whether the cited corpus is mentioned in its data, experiment, or results section. As can be seen in Table 1, corpora with fewer topics tend to be used more often in experiments than those with larger amounts. In total, 230 experiments were carried out on argument corpora with no clearly defined topic selection directive. The skew towards smaller-scale experiments may affect generalizability.

4 Bootstrapping The Argument Ontology

Topic ontologies provide for a knowledge organization principle, and, especially if widely accepted, also a standard. They are typically modeled as directed acyclic graphs, where nodes correspond to topics and edges indicate "is part of" relations. Topics that are part of other topics are called their subtopics. A topic ontology is often displayed in levels, starting with the topics that are not subtopics of others, continuing recursively with each lower level of subtopics. Figure 1 shows an excerpt of a two-level topic ontology for arguments.

The identification of the topics to be included in The Argument Ontology (TAO), as well as their relations, requires domain expertise. Building an all-encompassing ontology thus requires experts from every top-level domain where argumentation of scientific interest is expected. In the following, we suggest and outline three authoritative sources of expert topic ontologies, which comprise a wide selection of important argumentative topics. We use them to bootstrap a first version of TAO.

World Economic Forum (WEF) The World Economic Forum is a not-for-profit foundation that coordinates organizations from both the public and the private sector to work on economical and societal issues. As part of their efforts, their "Strategic Intelligence" platform[5] strives to inform decision makers on domestic and global topics, specifically global issues (e.g., artificial intelligence and climate change), industries (e.g., healthcare delivery and private investors), and economies (e.g., Africa and ASEAN). Domain experts for each topic curate a stream of relevant news articles which they each tag with 4-9 subtopics of their topic (e.g., the continuous monitoring of mental health).

Wikipedia Wikipedia strives for a neutral point of view, but many topics of public interest are discussed controversially. Some editors thus curate a list of controversial Wikipedia articles to highlight where special care is needed, grouped into 14 top-level topics (e.g., environment and philosophy) and 4-176 subtopics (e.g., creationism and pollution).[6] We omit the "People" topic and articles on countries; their controversiality is not universal.

Debatepedia The Debatepedia portal's goal is to create an encyclopedia of debates which are organized as "pro" and "con" arguments. A list of 89 topics helps visitors to browse the debates. The debates are contributed by anonymous web users. Topics in Debatepedia tend to address issues of Western culture. For example, the topic "United States" covers 306 debates while "Third World" covers only 12. The site is no longer maintained, but accessible through the Wayback Machine.[7]

The three ontologies are publicly accessible, and two of them are actively maintained and updated. Acquiring the ontologies is straightforward—not straightforward is to make use of them. A key task associated with every topic ontology is to categorize a given document. Having just a short string

[5] https://intelligence.weforum.org

[6] https://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues

[7] https://web.archive.org/web/20180222051626/http://www.debatepedia.org/en/index.php/Welcome_to_Debatepedia%21

Type	Count	Example topic		
		Topic label	Normalized form	Corpus
Concept	1,394	Abortion	abortion	Claim Sentence Search
Conclusion	707	We should ban partial birth abortion	partial birth abortion	Evidence Quality
Question	110	Should abortion be prohibited?	abortion	IAM
Imperative	25	Ban abortions	abortion	Record Debating Dataset 5
Comparison	23	Pro Choice vs. Pro Life	pro choice vs pro life	UKPConvArg1

Table 2: Counts of the topic types in the 39 preprocessed corpora with examples and their normalized form.

label describing a (potentially multifaceted) topic, such as "The Great Reset", renders this task exceedingly difficult. Fortunately, domain experts have been pre-categorizing documents into the aforementioned ontologies. In particular, regarding the WEF, invited domain experts categorize news articles for every topic, regarding Wikipedia, the text of the associated articles is available, as are the associated debates on Debatepedia.

Articles that are categorized into Level 2 topics are propagated up to their respective Level 1 topics. Table 3 shows the large differences between the ontologies. The WEF ontology contains the most topics, links the most documents, and has the most tokens overall. Wikipedia's Level 2 topics link to a single article each, yielding less text overall.

5 Topic Coverage

To assess the topic coverage of an argument corpus given the three ontologies, we map their topic labels (if provided) to matching ontology topics.

5.1 Topic Label Normalization

Table 1 lists 39 argument corpora that provide topic labels. Altogether 2,259 different labels have been assigned. They are concise descriptions of the main issues of an argument provided by the corpus authors. The labels possess the text register of the respective corpus: In essays, for instance, topics are usually thesis statements, while Wikipedia-derived corpora use article titles, and the topics of debate corpora include clichés such as "This house should". Often, topic labels express a stance towards a target issue, e.g., "ban guns". Five types of topic labels can be distinguished: concept, comparison of concepts, conclusion (includes claim and thesis), question, and imperative. We normalize the topic labels by converting all concepts to singular form, removing clichés, and dropping stance-indicating words such as "legalize". Our normalization aims at retaining only the central target issue of a topic label and leads to 798 unique topic labels.

5.2 Mapping Topic Labels to Ontology Topics

Using the preprocessed topic labels as queries, we retrieve for each topic label the 50 top-most relevant topics in each level of the three ontologies. To facilitate the retrieval of ontology topics, we employ a BM25-weighted (Robertson et al., 2004) index of the concatenated documents for each topic. This enables us to narrow down the mapping of a topic label to a manageable size. Except for a handful of cases, 50 ontology topics can be retrieved for each topic label. The topic labels were then manually mapped to an ontology topic, if they form synonyms, or if the former is a subtopic of the latter—which thus indicates that all arguments in the corpus with that topic label are about the ontology topic. A topic label can thus be mapped to multiple ontology topics. For example, the topic label "plastic bottles" is mapped to "pollution" and "recycling" in Wikipedia Level 2.

5.3 Analysis of Topic Coverage

Table 3 shows general statistics of this mapping of corpora topic labels to ontology topics. Most of the topic labels (2,141 out of 2,259) are mapped to at least one Debatepedia topic while only 395 labels are mapped to WEF Level 2 topics. For Wikipedia Level 2, only 298 out of the 748 topics are actually covered by argument corpora. This first analysis already suggests that existing argument corpora often only cover a small subset of possible argumentative topics that people are trained to debate. For those topic labels that can be mapped, they belong on average to 2.78 topics in Debatepedia, 1.24 topics in Wikipedia Level 1, and 1.53 topics in WEF Level 1. As discussed in Section 4, topics in Debatepedia focus on the Western culture and are easily accessible, whereas topics in WEF require in-depth domain knowledge and have more global relevance. The broad coverage of Debatepedia's topics indicates that argument corpora focus on common, widely discussed topics rather than global issues or those that need domain knowledge.

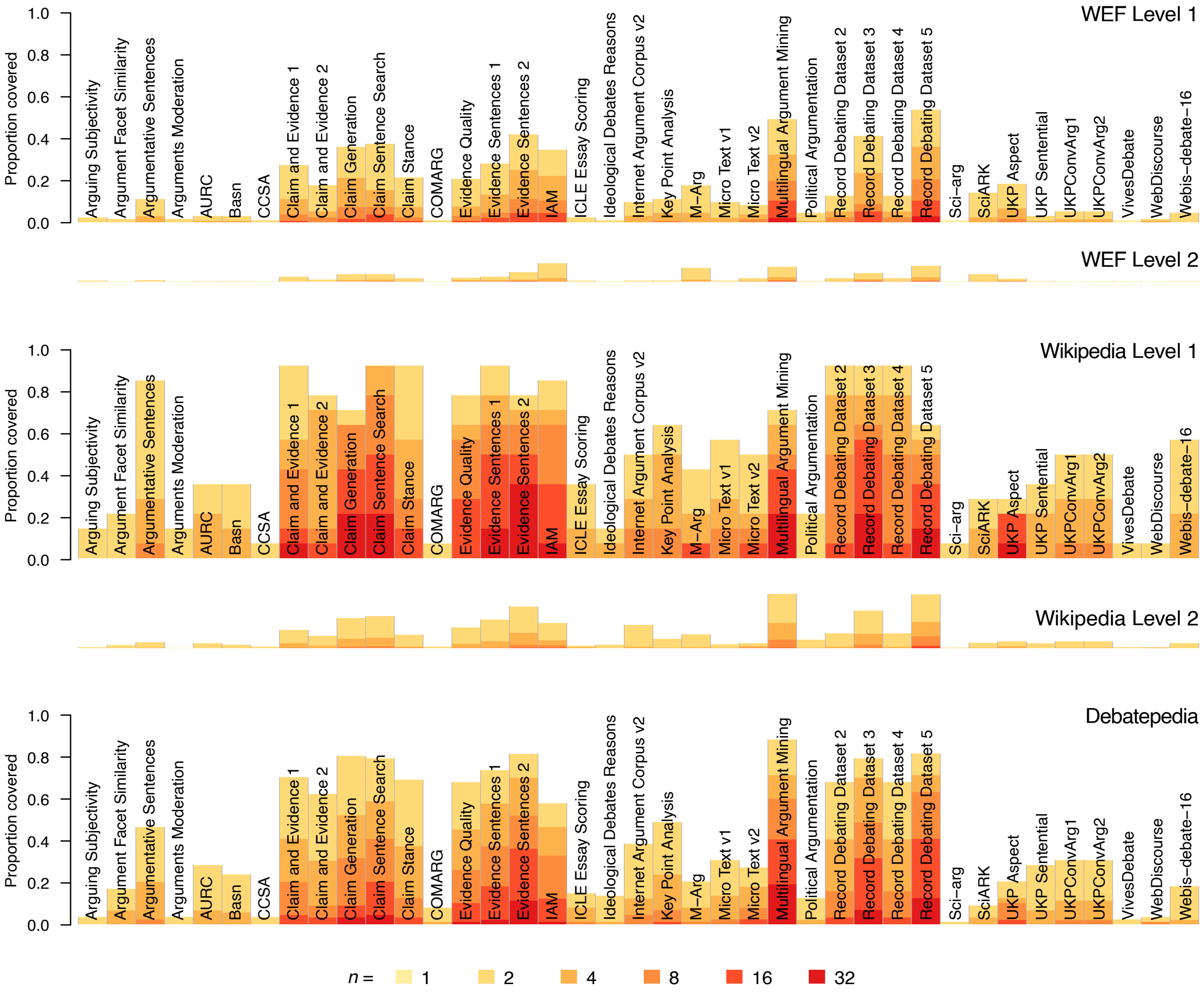


Figure 2: Proportion of ontology topics covered by at least n corpus topics (per ontology level and per corpus).

For a more fine-grained analysis, Figure 2 illustrates the differences regarding the number of ontology topics covered by a corpus: While topics in Wikipedia Level 1 are covered well by some argument corpora, topics in Wikipedia and WEF Level 2 are covered only marginally. Note that topic coverage varies significantly between the corpora: the Claim Sentence Search dataset's topics cover 93% of the Wikipedia Level 1 topics, while the Ideological Debates Reasons dataset covers only 14%. The colors show the topic granularity of the corpus; especially the Record Debating Dataset 3 dataset is fine-grained: as the highest value, 36 of its topics are mapped to the Wikipedia Level 1 category "Politics and Economics".

Figure 3 shows how the set of the units of the 39 labeled corpora distribute over the top-matching topics in Debatepedia, Wikipedia Level 1, and WEF Level 1. Distributions over Level 2 are omitted for brevity and can be found in Figure 4 in the Appendix. The distribution is significantly skewed: while the top ten topics in Debatepdia are matched by 354,811 to 138,407 corpora units, the top ten topics in WEF Level 1 are matched by 344,345 to 28,725 corpora units. This supports our finding that the corpora cover easily accessible topics (e.g., "Media and Entertainment" and "Society").

6 Unit Categorization

The previous analysis assesses argument corpora which contain topic labels. About a third of the argument corpora do not. As a heuristic step to assessing their topic coverage, we map the ontology topics for a unit (Table 1) in an argument corpus by treating the unit as a (long) query in a standard information retrieval setup, where ontology topics are the retrieval targets. The documents categorized into each topic have been concatenated and used as the topic's representation. Though the documents associated with a topic are not necessarily argumentative, they cover the salient topic aspects.

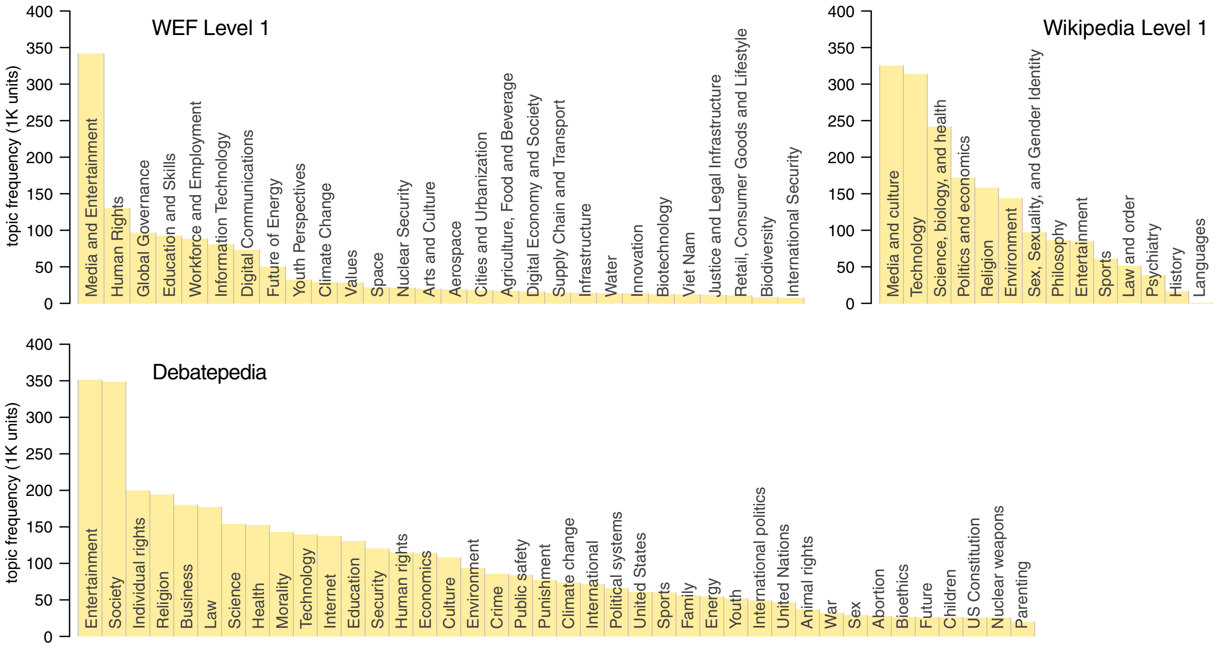


Figure 3: Distribution of corpora units over the top matching topics in an ontology (39 labeled corpora).

Ontology	**Acquired ontologies** (Section 4)				**Topic coverage** (Section 5)					**Unit categorization** (Section 6)										
	Topics	Topic statistics			Mapped topic labels	Covered ontology topics				Direct match			Semantic interpretation				Text2vec-SI			
		Authors	Docs	Tokens		All	Min	Mean	Max	P	R	F	Policy	P	R	F	Policy	P	R	F
WEF L1	137	334.1	940.7	490,576.6	1,339	92	1	1.53	13	0.38	0.23	0.29	$k = 12$	0.22	0.75	**0.34**	$k = 7$	0.19	0.53	0.28
WEF L2	822	216.8	550.3	310,229.7	395	154	1	1.56	22	0.59	0.11	0.19	$k = 30$	0.21	0.68	**0.32**	$\theta = 0.93$	0.15	0.49	0.23
WP L1	14	78,013.7	68.0	339,088.0	1,647	14	1	1.24	3	0.12	0.04	0.06	$k = 3$	0.32	0.65	0.43	$k = 2$	0.41	0.55	**0.47**
WP L2	748	1,929.5	1.0	6,149.1	1,560	298	1	1.80	16	0.47	0.34	0.40	$\theta = 0.05$	0.54	0.64	**0.59**	$\theta = 0.89$	0.22	0.52	0.31
DP	89	145.0	61.7	84,787.6	2,141	88	1	2.78	10	0.49	0.37	0.42	$\theta = 0.02$	0.52	0.61	**0.56**	$k = 23$	0.36	0.80	0.50

Table 3: Statistics for each topic ontology level: for topics and topic documents (Section 4), Count of mapped topic labels of the analyzed corpora for each ontology level, Count of all covered ontology topics by the topic labels and the min, max, and mean count of covered ontology topics per topic label (Section 5), and the effectiveness of the approaches and baseline in unit categorization (in terms of precision, recall, and F_1-score) (Section 6).

To retrieve topics for a corpus unit, we implement and evaluate the following approaches: Semantic Interpretation (SI) and SI with Text Embeddings (Text2vec-SI). The Semantic interpretation approach computes the semantic similarity of a unit and a topic as follows: it uses the cosine similarity of the TF-IDF vectors for the unit and the concatenated topic's documents. This corresponds to the semantic interpretation step that is at the core of the well-known ESA model (Gabrilovich and Markovitch, 2007). Text2vec-SI calculates the similarity of topics and corpus units using BERT embeddings (Devlin et al., 2018). Following common practice, we take the dimension-wise average of the word embeddings for all tokens in the text.[8] We tried other embeddings and approaches that performed similarly. The results of these approaches can be found in the appendix. As a baseline, we implement a direct match approach, which assigns a unit an ontology topic if the topic's text appears in the unit text (ignoring case).

For evaluation, we collect 34,638 pooled query relevance judgments (0.53 inter-annotator agreement as per Krippendorff's α) on 104 randomly selected argument units as queries from 26 corpora. The annotation process is detailed in the Appendix.

Based on the similarity scores of the approaches, we derive Boolean labels that indicate whether a unit is or is not about one of the ontologies' topics using two policies. The *threshold* policy labels a unit as about a topic if their similarity is above a threshold θ. The *top-k* policy labels a unit as about a topic if the topic is among the top-k topics with the highest similarity to the unit. We report the parameter of the policy that achieved the highest F_1-score on the pooled judgments for each approach.

[8] For efficiency, we limited the embeddings to 10,000 randomly sampled sentences for the topics that had more sentences associated with them.

Table 3 shows the results of this evaluation. The baseline produces different results across ontologies—it performs poorly for both the abstract topics in Wikipedia Level 1 and the specific topics in WEF Level 2. The semantic interpretation approach clearly outperforms the baseline for all ontologies in terms of the F_1-score. The Text2vec-SI approach outperforms the baseline and the semantic interpretation on abstract topics (Wikipedia Level 1), but its effectiveness is below that of the semantic interpretation approach on the other ontology levels.

7 Conclusion

The computational argumentation community risks topic bias in its approaches if the representativeness of topics in future corpora is not ensured. Achieving topic coverage is complicated by the fact that the landscape of controversial topics has not yet been well explored, and that there are no widely accepted ontologies for argument topics. In this paper, we venture into this future by mapping the landscape of argument topics and making it accessible for corpus construction and experimental design. We have identified three authoritative sources of ontological knowledge related to argument topics that provide an initial foundation for The Argument Ontology (TAO). For each source ontology, we evaluate the topic coverage of 39 argument corpora labeled with topics by matching the labels with the topics of the ontologies. To evaluate the topic coverage of corpora without topic labels, we develop an approach to identify the ontology topics of an argumentative text and achieve an F_1 of 0.59.

Our analyses show that the topic coverage of existing argument corpora is both limited to a subset of the topics of the ontologies and skewed. Most topics that require expertise, such as mental health, philosophy, or international security, are treated only peripherally in argumentation corpora. Therefore, existing argumentation technologies are more suited to teaching people how to construct arguments in general than to helping them make decisions about such and similarly complex topics. For the development of robust argumentation technologies, corpora need to be carefully drawn from a specific domain to allow for reliable experiments and the development of generalizable classifiers.

Future work for further development of TAO consists mainly of further surveying the argument topic landscape and unifying the various available ontologies. In addition to "is part of" relationships between topics, other relationship types can also be considered to build an argument topic knowledge base. However, our first version of TAO and our analyses can already help in selecting arguments for future corpus construction and model training.

Limitations

The three topic ontologies we used to evaluate topic coverage of argument corpora are from authoritative sources. Nevertheless, they probably do not cover all possible controversial topics relevant to argumentation (e.g., topics concerning private life). A comprehensive coverage of controversial topics in breadth and depth will likely remain an unattainable goal. Moreover, unifying the three thematic ontologies into a standard ontology is still an open problem given the many possible interpretations and relationships between the topics.

Another limitation is the moderate effectiveness achieved by our approaches for categorizing argument units. This is the case due to the large collection of controversial topics (about 748 for Wikipedia). Future research can be improved by using the structure of the topic ontology and hierarchical classifiers. Furthermore, it is also unclear whether the topic dependence of argumentation approaches decreases with increasing corpus size.

Ethics Statement

Our goal is to investigate whether and to what extent existing argumentation corpora are topic biased. This serves to critically examine the state of the art. However, we by no means want to give the impression that previous corpus authors lack ambition or diligence. Rather, the opposite is the case. The number of corpora that have been created in the last decade shows that the community is aware of the fact that not all areas of the argumentation landscape have been covered yet, and is therefore doing its utmost to explore it further. In a dynamic and rapidly growing research field, standards are usually developed in parallel with contributions, not in advance. Our research may therefore contribute to the further standardization of the corpus linguistics of argumentation.

The manual annotation of arguments and topics was done by expert annotators of our research groups. They were compensated fairly under German law. No personal data was collected.

References

Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn Walker. 2016. Internet Argument Corpus 2.0: An SQL Schema for Dialogic Social Media and the Corpora to go with it. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4445–4452. European Language Resources Association (ELRA).

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A Benchmark Dataset for Automatic Detection of Claims and Evidence in the Context of Controversial Topics. In *Proceedings of the 2014 Workshop on Argumentation Mining (ArgMining 2014)*, pages 64–68. Association for Computational Linguistics.

Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. 2019a. Modeling Frames in Argumentation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP 2019)*. ACL.

Yamen Ajjour, Henning Wachsmuth, Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. 2019b. Data Acquisition for Argument Search: The args.me corpus. In *42nd German Conference on Artificial Intelligence (KI 2019)*. Springer.

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein. 2016. Cross-Domain Mining of Argumentative Text through Distant Supervision. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2016)*, pages 1395–1404. Association for Computational Linguistics.

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein. 2016a. Cross-Domain Mining of Argumentative Text through Distant Supervision. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2016)*, pages 1395–1404. Association for Computational Linguistics.

Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016b. A News Editorial Corpus for Mining Argumentation Strategies. In *26th International Conference on Computational Linguistics (COLING 2016)*, pages 3433–3443. Association for Computational Linguistics.

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance Classification of Context-Dependent Claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, pages 251–261. Association for Computational Linguistics.

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020. From arguments to key points: Towards automatic argument summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2021)*, pages 4029–4039. Association for Computational Linguistics.

Roy Bar-Haim, Dalia krieger, Orith Toledo-Ronen, Lilach Edelstein, Yonatan Bilu, Alon Halfon, Yoav Katz, Amir Menczel, Ranit Aharonov, and Noam Slonim. 2019. From Surrogacy to Adoption; From Bitcoin to Cryptocurrency: Debate Topic Expansion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 977–990. Association for Computational Linguistics.

Floris Bex, John Lawrence, Mark Snaith, and Chris Reed. 2013. Implementing the Argument Web. *Communications of the ACM*, 56:66–73. Crawled in Jan, 2020.

Yonatan Bilu, Ariel Gera, Danel Hershcovich, Benjamin Sznajder, Dan Lahav, Guy Moshkowich, Anael Malet, Assaf Gavron, and Noam Slonim. 2019. Argument Invention from First Principles. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 1013–1026. Association for Computational Linguistics.

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

Tuhin Chakrabarty, Christopher Hidey, and Kathy McKeown. 2019. IMHO fine-tuning improves claim detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 558–563.

Liying Cheng, Lidong Bing, Ruidan He, Qian Yu, Yan Zhang, and Luo Si. 2022. Iam: A comprehensive and large-scale dataset for integrated argument mining tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 2277–2287.

Liying Cheng, Lidong Bing, Qian Yu, Wei Lu, and Luo Si. 2020. Ape: Argument pair extraction from peer review and rebuttal via multi-task learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7000–7011.

Alexander Conard, 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 (ExProM 2012)*, pages 80–88.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the Essence of a Claim? Cross-Domain Claim Identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 2045–2056. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Liat Ein-Dor, Eyal Shnarch, Lena Dankin, Alon Halfon, Benjamin Sznajder, Ariel Gera, Carlos Alzate, Martin Gleize, Leshem Choshen, Yufang Hou, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. 2020. Corpus wide argument mining - A working solution. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7683–7691.

Shnarch Eyal, Leshem Choshen, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. 2020. Unsupervised expressive rules provide explainability and assist human experts grasping new domains. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2678–2697. Association for Computational Linguistics.

Neele Falk, Iman Jundi, Eva Maria Vecchi, and Gabriella Lapesa. 2021. Predicting moderation of deliberative arguments: Is argument quality the key? In *Proceedings of the 8th Workshop on Argument Mining*, pages 133–141.

Aris Fergadis, Dimitris Pappas, Antonia Karamolegkou, and Harris Papageorgiou. 2021. Argumentation mining in scientific literature for sustainable development. In *Proceedings of the 8th Workshop on Argument Mining*, pages 100–111.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007*, pages 1606–1611.

Yingqiang Gao, Nianlong Gu, Jessica Lam, and Richard H.R. Hahnloser. 2022. Do discourse indicators reflect the main arguments in scientific papers? In *Proceedings of the 9th Workshop on Argument Mining*, pages 34–50.

Martin Gleize, Eyal Shnarch, Leshem Choshen, Lena Dankin, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. 2019. Are You Convinced? Choosing the More Convincing Evidence with a Siamese Network. In *Proceedings of the 2019 Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 967–976.

Alex Graves and Jürgen Schmidhuberab. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks : the official journal of the International Neural Network Society*, 18:602–10.

Shai Gretz, Yonatan Bilu, Edo Cohen-Karlik, and Noam Slonim. 2020. The workweek is the best time to start a family – a study of GPT-2 based claim generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 528–544, Online. Association for Computational Linguistics.

Thomas R. Gruber. 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5:199–220.

Giulia Grundler, Piera Santin, Andrea Galassi, Federico Galli, Francesco Godano, Francesca Lagioia, Elena Palmieri, Federico Ruggeri, Giovanni Sartor, and Paolo Torroni. 2022. Detecting arguments in cjeu decisions on fiscal state aid. In *Proceedings of the 9th Workshop on Argument Mining*, pages 143–157.

Yunfan Gu, Yhongyu Wei, Maoran Xu, Hao Fu, Yang Liu, and Xuanjing Huang. 2018. Incorporating Topic Aspects for Online Comment Convincingness Evaluation. In *Proceedings of the 5th Workshop on Argument Mining (ArgMining 2018)*, pages 97–104. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2016a. Argumentation Mining in User-Generated Web Discourse. *Computational Linguistics*.

Ivan Habernal and Iryna Gurevych. 2016b. What makes a convincing argument? Empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2016c. Which argument is more convincing? Analyzing and predicting convincingnessof Web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pages 1589–1599. Association for Computational Linguistics.

Shohreh Haddadan, Elena Cabrio, and Serena Villata. 2019. Yes, we can! Mining Arguments in 50 Years of US Presidential Campaign Debates. In *Proceedings of the 2019 Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 4684–4690.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are You Taking this Stance? Identifying and Classifying Reasons in Ideological Debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 751–762.

Annette Hautli-Janisz, Zlata Kikteva, Wassiliki Siskou, Kamila Gorska, Ray Becker, and Chris Reed. 2022. Qt30: A corpus of argument and conflict in broadcast debate. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3291–3300.

Birger Hjørland. 2001. Towards a theory of aboutness, subject, topicality, theme, domain, field, content . . . and relevance. *Journal of the American Society for Information Science and Technology*, 52(9):774–778.

Kialo. 2020. Kialo. www.kailo.com. Crawled in Jan, 2020.

Takahiro Kondo, Koki Washio, Katsuhiko Hayashi, and Yusuke Miyao. 2021. Bayesian argumentation-scheme networks: A probabilistic model of argument validity facilitated by argumentation schemes. In *Proceedings of the 8th Workshop on Argument Mining*, pages 112–124.

Anne Lauscher, Goran Glavaš, and Simone Paolo Ponzetto. 2018. An argument-annotated corpus of scientific publications. In *Proceedings of the 5th Workshop on Argument Mining*, pages 40–46. Association for Computational Linguistics.

Tamar Lavee, Matan Orbach, Lili Kotlerman, Yoav Kantor, Shai Gretz, Lena Dankin, Michal Jacovi, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. 2019. Towards Effective Rebuttal: Listening Comprehension using Corpus-Wide Claim Mining. In *Proceedings of the Fourth Workshop on Argument Mining 2017(ArgMining 2017)*, pages 719–724.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500.

Ran Levy, Ben Boginand Shai Gretz, Ranit Aharonov, and Noam Slonim. 2018. Towards an argumentative content search engine using weak supervision. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2066–2081.

Yinzi Li, Wei Chen, Zhongyu Wei, Yujun Huang, Chujun Wang, Siyuan Wang, Qi Zhang, Xuanjing Huang, and Libo Wu. 2022. A structure-aware argument encoder for literature discourse analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7093–7098.

Marco Lippi and Paolo Torroni. 2016. Argument Mining from Speech: Detecting Claims in Political Debates. In *Proceedings of the 2016 Association for the Advancement of ArtificialIntelligence (AAAI 2016)*, pages 2979–2985.

Siyi Liu, Sihao Chen, Xander Uyttendaele, and Dan Roth. 2021. Multioped: A corpus of multi-perspective news editorials. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4345–4361.

Stefano Menini, Elena Cabrio, Sara Tonelli, and SerenaVillata. 2018. Never Retreat, Never Retract: Argumentation Analysis for Political Speeches. In *Proceedings of the Thirty-second Association for the Advancement of Artifical Intelligene (AAAI) Conference of Artifical Intelligence*, pages 4889–4896. AAAI Press.

Rafael Mestre, Razvan Milicin, Stuart E. Middleton, Matt Ryan, Jiatong Zhu, and Timothy J Norman. 2021. M-arg: Multimodal argument mining dataset for political debates with audio and transcripts. In *Proceedings of the 8th Workshop on Argument Mining*, pages 78–88.

Shachar Mirkin, Guy Moshkowich, Matan Orbach, Lili Kotlerman, Yoav Kantor, Tamar Lavee, Michal Jacovi, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. 2018. Listening Comprehension over Argumentative Content. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 719–724.

Amita Misra, Brian Ecker, and Marilyn Walker. 2016. Measuring the similarity of sentential arguments in dialogue. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 276–287, Los Angeles. Association for Computational Linguistics.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic Detection of Arguments in Legal Texts. In *Proceedings of the 11th International conference on Artificial Intelligence and Law (ICAIL 2007)*, pages 225–230. Association for Computational Machinery.

Lily Ng, Anne Lauscher, Joel Tetreault, and Courtney Napoles. 2020. Creating a domain-diverse corpus for theory-based argument quality assessment. In *Proceedings of the 7th Workshop on Argument Mining*, pages 117–126, Online. Association for Computational Linguistics.

Matan Orbach, Yonatan Bilu, Ariel Gera, Yoav Kantor, Lena Dankin, Tamar Lavee, Lili Kotlerman, Shachar Mirkin, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. A dataset of general-purpose rebuttal. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5591–5601, Hong Kong, China. Association for Computational Linguistics.

Matan Orbach, Yonatan Bilu, Assaf Toledo, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2020. Out of the echo chamber: Detecting countering debate speeches. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7073–7086, Online. Association for Computational Linguistics.

Joonsuk Park and Claire Cardie. 2018. A Corpus of e-Rulemaking User Comments for Measuring Evaluability of Arguments. In *Proceedings of the 2018 International Conference on Language Resources and Evaluation (LREC 2018)*.

Andreas Peldszus and Manfred Stede. 2015. An annotated corpus of argumentative microtexts. In *Proceedings of the 2015 European Conference on Argumentation: Argumentation and Reasoned Action (ECA 2015)*.

Isaac Persing, Alan Davis, and Vincent Ng. 2010. Modeling Organization in Student Essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 229–239.

Isaac Persing and Vincent Ng. 2017. Lightly-Supervised Modeling of Argument Persuasiveness. In *Proceedings of 2017 International Joint Conference on Natural Language Processing (IJCNLP 2017)*, pages 594–604.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, pages 2227–2237. ACL.

Prakash Poudyal, Jaromir Savelka, Aagje Ieven, Marie Francine Moens, Teresa Goncalves, and Paulo Quaresma. 2020. ECHR: Legal corpus for argument mining. In *Proceedings of the 7th Workshop on Argument Mining*, pages 67–75, Online. Association for Computational Linguistics.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and Clustering of Arguments with Contextualized Word Embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 567–578. Association for Computational Linguistics.

Myrthe Reuver, Suzan Verberne, Roser Morante, and Antske Fokkens. 2021. Is Stance Detection Topic-Independent and Cross-topic Generalizable? – A Reproduction Study. In *Proceedings of the 2021 Workshop on Argumentation Mining (ArgMining 2021)*.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show Me Your Evidence - an Automatic Method for Context Dependent Evidence Detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, pages 719–724.

Stephen E. Robertson, Hugo Zaragoza, and Michael J. Taylor. 2004. Simple BM25 extension to multiple weighted fields. In *Proceedings of the 2004 ACM CIKM International Conference on Information and Knowledge Management, Washington, DC, USA, November 8-13, 2004*, pages 42–49. ACM.

Allen Roush and Arvind Balaji. 2020. Debatesum: A large-scale argument mining and summarization dataset. In *Proceedings of the 7th Workshop on Argument Mining*, pages 1–7, Online. Association for Computational Linguistics.

Ramon Ruiz-Dolz, Montserrat Nofre, Mariona Taulé, Stella Heras, and Ana García-Fornes. 2021. Vivesdebate: A new annotated multilingual corpus of argumentation in a debate tournament. *Applied Sciences*, 11(15):7160.

Eyal Schnarch, Carlos Alzate, Lena Dankin, Martin Gleize, Yufang Hou, Leshem Choshen, Ranit Aharonov, and Noam Slonim. 2018. Will it Blend? Blending Weak and Strong Labeled Data in a Neural Network for Argumentation Mining. In *Proceedings of the 2018 Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pages 599–605.

Carlos Silla and Alex Freitas. 2011. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22:31–72.

Maria Skeppstedt, Andreas Peldszus, and ManfredS Stede. 2018. More or less controlled elicitation of argumentative text: Enlarging a microtext corpus via crowdsourcing. In *Proceedings of the 5th Workshop on Argument Mining 2017 (ArgMining 2017)*, pages 155–163. Association for Computational Linguistics.

Yangqiu Song and Dan Roth. 2015. Unsupervised sparse vector densification for short text similarity. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 1275–1280.

Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. 2018. ArgumenText: Searching for Arguments in Heterogeneous Sources. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 21–25.

Christian Stab and Iryna Gurevych. 2017. Parsing Argumentation Structure in Persuasive Essays. *Computational Linguistics*, 43(3):619–659.

Aixin Sun and Ee-Pen Lim. 2001. Hierarchical Text Classification and Evaluation. In *Proceedings of the 2001 Institute of Electrical and Electronics Engineer (IEEE) International Conference on Data Mining (ICDM 2001)*, pages 521–528. Association for Computational Linguistics.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web(WWW 2016)*, pages 613–624. International World Wide Web Conferences Steering Committee.

Orith Toledo-Ronen, Matan Orbach, Yonatan Bilu, Artem Spector, and Noam Slonim. 2020. Multilingual argument mining: Datasets and analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 303–317. Association for Computational Linguistics.

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

Dietrich Trautmann, Johannes Daxenberger, Christian Stab, Hinrich Schütze, and Iryna Gurevych. 2020. Fine-grained argument unit recognition and classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9048–9056.

Frans H. van Eemeren, editor. 2015. *Reasonableness and Effectiveness in Argumentative Discourse*, volume 27 of *Argumentation Library*. Springer.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational Argumentation Quality Assessment in Natural Language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EMNLP 2017)*, pages 176–187.

Eduardo Xamena, Nélida Beatriz Brignole, and Ana Gabriela Maguitman. 2017. A structural analysis of topic ontologies. *Information Science*, 421:15–29.

Stephen Yablo. 2014. *Aboutness*. Princeton University Press.

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. Conversational Flow in Oxford-style Debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2016)*.

A Appendix

A.1 Mapping Topic Labels to Level 2 Topics

For completeness, Figure 4 shows the two graphs that are omitted from Figure 3 of the paper as their fine-grained topics are less relevant for the discussion in Section 5.3.

A.2 Annotation Procedure for Unit Categorization

In order to assess the effectiveness of the approaches and baseline outlined in the paper, we employ a pooled evaluation, as it is standard for information retrieval evaluations, where there are too many instances for a complete manual annotation. We randomly sampled four units from 26 corpora, which were all annotated by three expert annotators. The annotators were instructed to label a topic as about the unit if they could imagine a discussion on the topic for which the unit would be relevant. For each unit, we annotated for aboutness only those topics which are among the five topics with the highest similarity to this unit according to at least one of the approaches. The employed assessment interface (see Figure 5) shows the unit (top left), the current topic (top right), as well as all topics in the pool for that unit (bottom; the current topic is marked blue, whereas already annotated topics are marked green (about) and red (not about). The same interface has been used for the topic label annotations.

To reduce biases, both the units and the topics were shown in a different and random order to each assessor. The annotation took about 40 hours. The annotation process resulted in an inter-annotator agreement of 0.53 in terms of Krippendorff's α and produced a total of 34,638 annotations of topic-unit pairs, about 2% of what would have been needed for a complete annotation.

A.3 Additional Unit Categorization Approaches

In addition to the approaches listed in Section 6 we used additional approaches and a baseline which we list here. The additional baseline randomly classifies corpora units as per the prior topic probability of each ontology level.

SI with Word Embeddings (W2V-SI) Adapted from Dense-ESA (Song and Roth, 2015), this approach represents each token by its TFIDF-weighted Word2vec embedding vector, and uses the highest cosine similarity between two vectors, one from each text, as the semantic similarity. To limit this quadratic effort, we use only the 100 tokens of each text with the highest TFIDF-score.

Text2vec-SI As a variant for BERT (Devlin et al., 2018), we embedded the ontology documents and corpora units using ELMo (Peters et al., 2018).

(a) Debatepedia

Topic	Covering units
Islam and the West	50
Islam	50
2008/2009 economic crisis	66
European Union	93
Middle East	134
Prison	180
China	254
Terrorism	521
Latin America	528
HIV/AIDS	579
Church and state	654
US legislation	845
Corruption	1,065
Welfare	1,369
Africa	1,435
Israeli-Palestinian conflict	1,541
Life and death	1,598
Languages	1,808
Privacy	2,312
Bush administration	2,702
Iraq	2,720
Weapons proliferation	2,790
Third world	3,208
Taxes	3,562
Disease	3,619
Obama administration	3,765
Conflict	4,950
Asia	5,016
Immigration	5,170
Race	6,086

(b) Wikipedia Level 2

Topic	Covering units
Irredentism	2
American Civil Liberties Union	2
Hezbollah	2
Esports	3
Separation of church and state	4
Birth defect	5
Quebec	5
Rape	6
Hurricane Katrina	6
Crime in the United States	6
Sexual abuse	6
Sex offender	6
Pacifism	7
Cyberstalking	7
Brexit	9
Economy of Japan	12
USA PATRIOT Act	12
Playboy Magazine	15
Super Bowl XXXVIII	19
Sexual harassment	20
Media bias	29
Culture war	35
Hip hop culture	35
European culture	35
Anime	40
East Germany	46
Communist state	46
Communist Party of China	46
Communist government	46
Communism	46

(c) World Economic Forum Level 1

Topic	Covering units
Agile Governance	1
Institutional Investors	1
Digital Identity	7
United Kingdom	9
Mexico	12
Behavioural Sciences	15
Canada	26
Corruption	31
Illicit Economy	54
Future of Economic Progress	66
Forests	74
European Union	93
Real Estate	132
Insurance and Asset Management	142
Humanitarian Action	232
3D Printing	254
China	258
Drones	260
Internet Governance	268
Cybersecurity	298
Internet of Things	320
Precision Medicine	339
Oceans	345
Latin America	516
Financial and Monetary Systems	608
Arctic	614
Banking and Capital Markets	634
Mining and Metals	656
Public Finance and Social Protection	778
Middle East and North Africa	928

(d) World Economic Forum Level 2

Topic	Covering units
Healthcare Human Capital	2
Environmentally-Sustainable Consumerism	2
Sustainable Consumption	3
Aquaculture	4
Urbanization and Circular Practices	5
Accelerating Sustainability	5
Forest Landscape Restoration	5
Stabilizing Economies, Keeping Protections	10
The Social Cost of Carbon	13
The Trump Presidency	20
New Leadership	20
Canada and Sustainable Energy	21
Economic Institutions	34
Outbound and Long-Term Investment	34
Deepening Interdependence	34
Digital Trade	34
Geopolitical and Geo-economic Recalibration	34
Pricing Climate into Finance	34
Trade and Investment	34
Trade and the Environment	34
Transnational Actors	34
Economic Integration	34
Healthcare Technology	47
Geo-strategic Competition	54
Energy-Related Emission Reduction	61
Energy Finance and Investment	61
Energy Access	61
Environmental Footprint	61
Electricity Decentralization	61
Electricity System Integration	61

Table 4: For each ontology except Wikipedia Level 1 the 30 topics with the least (but at least 1) units from the argument corpora covering them. All 14 topics of Wikipedia Level 1 are covered well and thus omitted here.

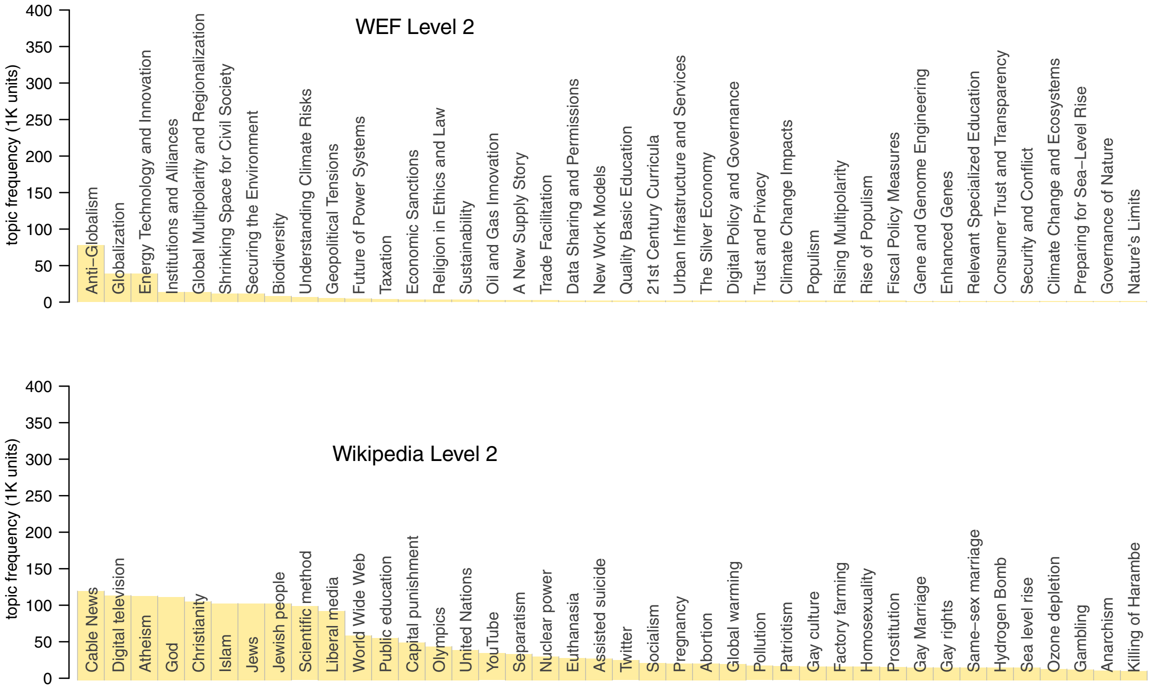
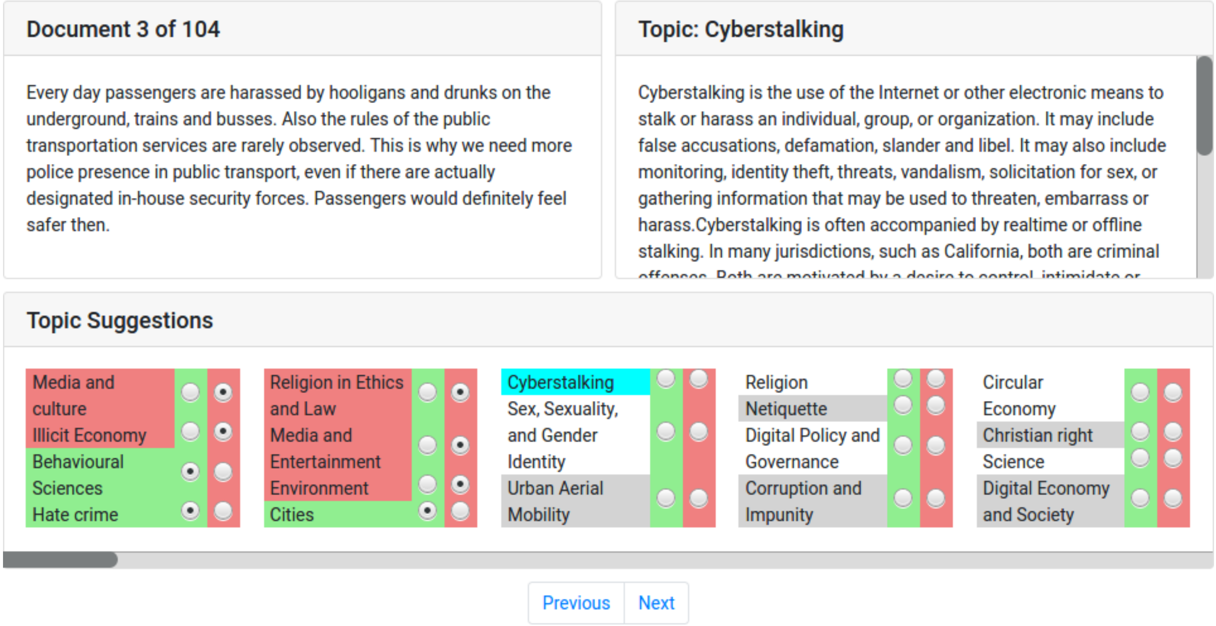


Figure 4: Omitted graphs from Figure 3, Section 5.3

Table 5 lists the results of all approaches for all thresholds and ranks.



Document 3 of 104

Every day passengers are harassed by hooligans and drunks on the underground, trains and busses. Also the rules of the public transportation services are rarely observed. This is why we need more police presence in public transport, even if there are actually designated in-house security forces. Passengers would definitely feel safer then.

Topic: Cyberstalking

Cyberstalking is the use of the Internet or other electronic means to stalk or harass an individual, group, or organization. It may include false accusations, defamation, slander and libel. It may also include monitoring, identity theft, threats, vandalism, solicitation for sex, or gathering information that may be used to threaten, embarrass or harass.Cyberstalking is often accompanied by realtime or offline stalking. In many jurisdictions, such as California, both are criminal offences. Both are motivated by a desire to control, intimidate or

Topic Suggestions

Media and culture	○	●
Illicit Economy	○	●
Behavioural Sciences	●	○
Hate crime	●	○
Religion in Ethics and Law	○	●
Media and Entertainment	○	●
Environment	○	●
Cities	●	○
Cyberstalking	○	○
Sex, Sexuality, and Gender Identity	○	○
Urban Aerial Mobility	○	○
Religion	○	○
Netiquette	○	○
Digital Policy and Governance	○	○
Corruption and Impunity	○	○
Circular Economy	○	○
Christian right	○	○
Science	○	○
Digital Economy and Society	○	○

Figure 5: Assessment interface for topic labeling.

Approach	**World Economic Forum**								**Wikipedia**								**Debatepedia**			
	Level 1				Level 2				Level 1				Level 2							
Baselines		P	R	F		P	R	F		P	R	F		P	R	F		P	R	F
Random		0.02	0.02	0.02		0.01	0.01	0.01		0.11	0.11	0.11		0.01	0.01	0.01		0.07	0.07	0.07
Direct match		0.38	0.23	0.29		0.59	0.11	0.19		0.12	0.04	0.06		0.47	0.34	0.40		0.49	0.37	0.42
By threshold	θ	P	R	F	θ	P	R	F	θ	P	R	F	θ	P	R	F	θ	P	R	F
Semantic interpretation	0.02	0.18	0.63	0.28	0.02	0.20	0.58	0.29	0.01	0.29	0.51	0.37	0.05	0.54	0.64	**0.59**	0.02	0.52	0.61	**0.56**
W2V-SI	0.20	0.14	0.63	0.23	0.16	0.13	0.63	0.21	0.10	0.12	0.55	0.20	0.11	0.18	0.69	0.29	0.07	0.29	0.81	0.43
Text2vec-SI$_{\mathrm{ELMo}}$	0.87	0.18	0.32	0.23	0.80	0.13	0.65	0.22	0.74	0.23	0.45	0.31	0.76	0.25	0.45	0.32	0.87	0.47	0.46	0.47
Text2vec-SI$_{\mathrm{BERT}}$	0.94	0.19	0.39	0.25	0.93	0.15	0.49	0.23	0.92	0.36	0.22	0.27	0.89	0.22	0.52	0.31	0.92	0.36	0.64	0.46
By rank	k	P	R	F	k	P	R	F	k	P	R	F	k	P	R	F	k	P	R	F
Semantic interpretation	12	0.22	0.75	**0.34**	30	0.21	0.68	**0.32**	3	0.32	0.65	0.43	12	0.39	0.70	0.50	19	0.43	0.78	**0.56**
W2V-SI	83	0.13	0.94	0.24	439	0.12	0.77	0.21	14	0.11	1.00	0.20	290	0.16	0.77	0.27	61	0.27	0.86	0.42
Text2vec-SI$_{\mathrm{ELMo}}$	4	0.25	0.44	0.32	42	0.13	0.68	0.23	2	0.39	0.53	0.45	46	0.18	0.64	0.28	13	0.43	0.71	0.54
Text2vec-SI$_{\mathrm{BERT}}$	7	0.19	0.53	0.28	80	0.11	0.66	0.20	2	0.41	0.55	**0.47**	80	0.17	0.70	0.28	23	0.36	0.80	0.50

Table 5: Performance of semantic interpretation approaches in human evaluation for each topic ontology level in terms of precision (P), recall (R), and F$_1$-score (F) for the "aboutness" label. For methods other than the baselines the table shows the values for both the similarity threshold θ and rank k that lead to the highest F$_1$-score respectively. The best F$_1$-scores for each ontology level are marked bold.