Addressing Controversial Topics in Search Engines

From the
Faculty of Media
of the
Bauhaus-Universität Weimar
Germany

The Submitted Dissertation of
Yamen Ajjour

To Obtain the Academic Degree of
Dr. rer. nat.

Advisor:	Prof. Dr. Benno M. Stein
Reviewer:	Prof. Dr. Chris Biemann
Oral Exam:	June 8, 2023

Contents

Preface

Abstract

Search engines are very good at answering queries that look for facts. Still, information needs that concern forming opinions on a controversial topic or making a decision remain a challenge for search engines. Since they are optimized to retrieve satisfying answers, search engines might emphasize a specific stance on a controversial topic in their ranking, amplifying bias in society in an undesired way. Argument retrieval systems support users in forming opinions about controversial topics by retrieving arguments for a given query. In this thesis, we address challenges in argument retrieval systems that concern integrating them in search engines, developing generalizable argument mining approaches, and enabling frame-guided delivery of arguments.

Adapting argument retrieval systems to search engines should start by identifying and analyzing information needs that look for arguments. To identify questions that look for arguments we develop a two-step annotation scheme that first identifies whether the context of a question is controversial, and if so, assigns it one of several question types: factual, method, and argumentative. Using this annotation scheme, we create a question dataset from the logs of a major search engine and use it to analyze the characteristics of argumentative questions. The analysis shows that the proportion of argumentative questions on controversial topics is substantial and that they mainly ask for reasons and predictions. The dataset is further used to develop a classifier to uniquely map questions to the question types, reaching a convincing F1-score of 0.78.

While the web offers an invaluable source of argumentative content to respond to argumentative questions, it is characterized by multiple genres (e.g., news articles and social fora). Exploiting the web as a source of arguments relies on developing argument mining approaches that generalize over genre. To this end, we approach the problem of how to extract argument units in a genre-robust way. Our experiments on argument unit segmentation show that transfer across genres is rather hard to achieve using existing sequence-to-sequence models.

Another property of text which argument mining approaches should generalize over is topic. Since new topics appear daily on which argument mining approaches are not trained, argument mining approaches should be developed in a

topic-generalizable way. Towards this goal, we analyze the coverage of 31 argument corpora across topics using three topic ontologies. The analysis shows that the topics covered by existing argument corpora are biased toward a small subset of easily accessible controversial topics, hinting at the inability of existing approaches to generalize across topics. In addition to corpus construction standards, fostering topic generalizability requires a careful formulation of argument mining tasks. Same side stance classification is a reformulation of stance classification that makes it less dependent on the topic. First experiments on this task show promising results in generalizing across topics.

To be effective at persuading their audience, users of an argument retrieval system should select arguments from the retrieved results based on what frame they emphasize of a controversial topic. An open challenge is to develop an approach to identify the frames of an argument. To this end, we define a frame as a subset of arguments that share an aspect. We operationalize this model via an approach that identifies and removes the topic of arguments before clustering them into frames. We evaluate the approach on a dataset that covers 12,326 frames and show that identifying the topic of an argument and removing it helps to identify its frames.

Abstract (in German)

Suchmaschinen sind sehr gut darin, Suchanfragen zu beantworten, die nach Fakten suchen. Dennoch bleibt der Informationsbedarf, der die Meinungsbildung zu einem kontroversen Thema oder die Entscheidungsfindung betrifft, eine Herausforderung für Suchmaschinen. Da sie darauf optimiert sind, befriedigende Antworten zu liefern, könnten Suchmaschinen in ihrem Ranking eine bestimmte Haltung zu einem kontroversen Thema hervorheben und damit Verzerrungen in der Gesellschaft in unerwünschter Weise verstärken. Argument-Retrieval-Systeme unterstützen Benutzer bei der Meinungsbildung zu kontroversen Themen, indem sie Argumente für eine bestimmte Suchanfrage abrufen. In dieser Arbeit befassen wir uns mit Herausforderungen in Argument-Retrieval-Systemen, die ihre Einbindung in Suchmaschinen, die Entwicklung generalisierbarer Argument-Mining-Ansätze und die Ermöglichung einer Bereitstellung von Argumenten zusammen mit deren Deutungsrahmen betreffen.

Die Einbindung von Argument-Retrieval-Systeme in Suchmaschinen sollte mit der Identifizierung und Analyse von Informationsbedürfnissen beginnen, die nach Argumenten suchen. Um Fragen zu identifizieren, die nach Argumenten suchen, entwickeln wir ein zweistufiges Annotationsschema, das erst feststellt, ob der Kontext einer Frage kontrovers ist, und wenn ja, sie einem von mehreren Fragetypen zuordnet: sachlich, methodisch und argumentativ. Anhand dieses Annotationsschemas erstellen wir einen Fragedatensatz aus den Query-Logs einer bedeutenden Suchmaschine in Russland und verwenden ihn, um die Merkmale argumentativer Fragen zu analysieren. Die Analyse zeigt, dass der Anteil an argumentativen Fragen zu kontroversen Themen hoch ist und dass sie hauptsächlich nach Gründen und Vorhersagen fragen. Der Datensatz wird außerdem verwendet, um einen Klassifikator zu entwickeln, der Fragen eindeutig den Fragetypen zuordnet und einen überzeugenden F1-score von 0,78 erreicht.

Das Web bietet zwar eine unschätzbare Quelle für argumentative Inhalte zur Beantwortung argumentativer Fragen, ist aber durch verschiedene Genres geprägt (z.B. Nachrichtenartikel und soziale Foren). Um das Web als Quelle für Argumente zu nutzen, müssen Argument-Mining-Ansätze entwickelt werden, die sich über verschiedene Genres generalisieren lassen. Zu diesem Zweck nähern wir uns dem Problem der Extraktion von Argumentationseinheiten auf eine genrerobuste Weise. Unsere Experimente zur Segmentierung von Argumenteinheiten zeigen, dass eine Übertragung über Genres mit bestehenden Sequenz-zu-Sequenz-Modellen nur schwer zu erreichen ist.

Eine weitere Eigenschaft von Text, die darüber Argument-Mining-Ansätze generalisieren sollten, ist das Thema. Da täglich neue Themen auftauchen, auf die Argument-Mining-Ansätze nicht trainiert sind, sollten Argument-Mining-Ansätze auf eine themengeneralisierbare Weise entwickelt werden. Um dieses Ziel zu erreichen, analysieren wir die Abdeckung von 31 Argumentkorpora über die Themen

von drei Themenontologien. Die Analyse zeigt, dass die Themen, die von den existierenden Argumentkorpora abgedeckt werden, auf eine kleine Teilmenge von leicht zugänglichen kontroversen Themen ausgerichtet sind, was auf die Unfähigkeit der existierenden Argument-Mining-Ansätze hinweist, über Themen zu generalisieren. Außer den Standards für die Korpuskonstruktion erfordert die Förderung der Generalisierbarkeit über Themen eine sorgfältige Formulierung der Aufgaben für das Argument Mining. Same side stance classification ist eine Neuformulierung der Haltungsklassifizierung, die sie weniger abhängig vom Thema macht. Erste Experimente zu dieser Aufgabe zeigen vielversprechende Ergebnisse bei der Generalisierung über Themen.

Um ihre Zuhörer zu überzeugen, sollten die Benutzer eines Argument-Retrieval-Systems aus den Ergebnissen Argumente auswählen, die bestimmte Deutungsrahmen eines kontroversen Themas betonen. Eine Herausforderung besteht darin, einen Ansatz zu entwickeln, mit dem die Deutungsrahmen eines Arguments identifiziert werden können. Zu diesem Zweck definieren wir einen Deutungsrahmen als eine Teilmenge von Argumenten, die einen Aspekt gemeinsam haben. Wir operationalisieren dieses Modell durch einen Ansatz, der das Thema von Argumenten identifiziert und entfernt, bevor er sie in Deutungsrahmen clustert. Wir evaluieren den Ansatz anhand eines Datensatzes, der 12.326 Argumente umfasst, und zeigen, dass die Identifizierung des Themas eines Arguments und dessen Entfernung zur Identifizierung seiner Deutungsrahmen beiträgt.

Ehrenwörtliche Erklärung

Ich erkläre hiermit ehrenwörtlich, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Teile der Arbeit, die bereits Gegenstand von Prüfungsarbeiten waren, sind ebenfalls unmissverständlich gekennzeichnet.

Weitere Personen waren an der inhaltlich-materiellen Erstellung der vorliegenden Arbeit nicht beteiligt. Insbesondere habe ich hierfür nicht die entgeltliche Hilfe von Vermittlungs- bzw. Beratungsdiensten (Promotionsberater oder anderer Personen) in Anspruch genommen. Niemand hat von mir unmittelbar oder mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen. Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt. Ich versichere, dass ich nach bestem Wissen die reine Wahrheit gesagt und nichts verschwiegen habe.

Weimar, 21. Mai 2022

Yamen Ajjour

Chapter 1

Introduction

The web is the largest source of information that has ever existed, allowing access to up-to-date content wherever needed. The fact that anybody can contribute and edit the content on the web makes it an open source of information that fosters diversity and collaboration. However, the open nature of the web also makes it a possible source for misinformation, one-sided documents, and fake news.

The omnipresence of digital media established search engines as a primary tool for filtering information. Search engines sort billions of documents in milliseconds to find relevant answers to a query. Search engines excel most at information needs that look for short answers that are widely acceptable (factual questions).

However, search engines struggle at addressing information needs that are related to forming an opinion on a topic. Such situations arise while making a decision on a private matter (e.g., "Should I buy or rent?") or when contending with another person on a particular policy (e.g., "Should we ban plastic bottles?"). Since search engines are optimized to retrieve satisfying answers in their top results, search engines might accidentally emphasize a certain view on the topic. In this way, search engines might deprive users of their right to reason about the different opinions on the topic to form their own.

Argumentation is the communicative activity of forming arguments and exchanging them to resolve a disagreement of opinions [168]. An argumentation is distinguished by a topic that defines what the disagreement is about. We can differentiate between two types of argumentation: dialogue and monologue. A monologue is a speech given by one person to an audience without an immediate contention. In a dialogue, two or more sides interactively participate in a discussion to support their stance on the topic.

Finding good arguments has been studied since the time of Aristotle, who introduced a theory on the art of persuasion (Rhetoric). According to this theory, an effective speech is characterized by three aspects: establishing the authority of the speaker (ethos), proper usage of logic (logos), and appealing to the emotions of the audience (pathos) [14]. Understanding the audience of an argument and selecting arguments that fit their background is an essential cornerstone of the pragma-dialectics view on argumentation [168].

Modeling an argument to represent its elements and the context in which it is exchanged is crucial to assess how good it is. Toulmin [162] introduced a formal argument model that defines an argument as a combination of six elements, among which are claim, data, and rebuttal. The claim of an argument is a proposition that is put forward for general acceptance. Data is established evidence on which the claim is grounded. A rebuttal is a counterargument that challenges the argument.

Computational argumentation is a research area that aims at modeling arguments in natural text to help users find and construct good arguments. Researchers develop methods to extract arguments from natural text, classify their stances into pro and con, and evaluate how good they are. The methods used in computational argumentation are developed based on corpora. A corpus is a set of natural texts that are segmented into units and labeled with their argumentative roles (e.g., claim or premise), stance (pro or con), or how persuasive they are. Corpora are usually collected starting from a set of pre-defined controversial topics (e.g., "legalize marijuana"). Controversial topics are those topics that generate strong disagreement among large groups of people [77].

Argument retrieval is an application of computational argumentation that aims at retrieving pro and con arguments for a given query. An essential goal of argument retrieval is helping users to effectively persuade an audience with a stance on a given controversial topic. An argument retrieval system relies on subsystems to extract arguments from a document source. Then, the arguments are categorized by their stances into pro and con. Finally, the system ranks arguments according to their relevance and quality.

This thesis aims at developing methods that allow:

1. Integrating argument retrieval systems in web search engines by identifying and analyzing argumentative questions.

2. Providing argument retrieval systems with an up-to-date and comprehensive source of arguments by developing generalizable argument mining methods.

3. Helping users to find arguments that suit their audience by identifying the frames of an argument.

1.1 Contributions and Research Questions

1.1.1 Identifying and Analyzing Argumentative Questions

Web search queries related to forming an opinion on a subject matter should be answered with all existing perspectives on it. Argument retrieval systems answer such queries with arguments that are categorized into pro and con. Retrieving arguments lefts the burden of taking a specific stance on the query from the shoulder of the search engine and keeps this for the user. Also, retrieving arguments explicates the reasoning structure either locally by showing the conclusion and premises of an argument or globally by retrieving supporting or attacking arguments to a given query.

While argument retrieval systems show a promising application in the context of web search, an understanding of what queries should be answered with arguments is still lacking. Research on argument retrieval assumes a query to be a controversial topic (e.g., "abortion") or a claim ("Abortion should be banned.") In web search, however, a query on a controversial topic might look for facts and not necessarily arguments.

The abundance of automated personal assistants is shaping information needs more in a conversational manner. Instead of typing queries, users are formulating questions in a more natural manner while looking up information. Research shows that queries are more and more being issued to search engines as questions [116]. This makes question-like queries a prominent element in the query stream of search engines. Hence, we raise the research question

(RQ1): *How to identify argumentative questions in the context of search engines?*

Argumentative questions are those that look for arguments or opinions for a given topic. To answer this research question, we prepare a question dataset that targets 19 controversial topics. The question dataset is sampled from the query log of Yandex for the year 2012. Each question is first labeled with whether it targets a given controversial topic or not. Those questions that are about a controversial topic are then annotated using a question taxonomy that is tailored to controversial topics. The question taxonomy covers three question types: *factual*, *method*, and *argumentative*. An analysis of the created dataset shows that the proportion of argumentative questions on controversial topics is high, reaching 28% of all questions. Using the dataset, we analyze the characteristics of argumentative questions and compare them with that of the other question types. The analysis shows that looking for reasons and predictions are among the most distinguishing features of argumentative questions.

Enabling an automatic integration of argument retrieval systems in the context of web search requires developing a method to identify those questions that are argumentative. To this end, we conduct experiments on the dataset to develop a supervised classifier that categorizes a given question into the three types: factual, method, and argumentative. The experiments are conducted in a leave-one-topic-out fashion, i.e., the test dataset covers one controversial topic while the training dataset covers the remaining 19 controversial topics. The experiments show that transformer-based classifiers can classify the three question types with an F1-score of 0.78. The results show the feasibility of identifying argumentative questions in the query stream of a search engine.

1.1.2 Generalizability of Argument Mining Approaches

The web offers an invaluable source of argumentative content that enables search engines to respond to argumentative questions. Existing argument mining approaches rely heavily on the web to develop computational models for argumentation. However, existing argument mining approaches are developed for single genres. For example, Stab and Gurevych [150] crawled 402 student essays from a writing support forum and used them to develop approaches to extract arguments. A genre is a set of texts that have a common linguistic function (e.g., editorials or argumentative essays).

Argument Mining is the task of extracting the argumentation structure from natural text. Researchers annotate the documents of a corpus with arguments based on a given argument model. An argument is a set of argument units and their relations. An argument model is an abstraction from the language level to a more formal level, where the constituting units of an argument, their relations, and their types are described. Using these corpora, researchers develop approaches to extract arguments that range from supervised classifiers to simple heuristics that utilize the structure of documents (e.g., categorized pro and con lists). A crucial step in argument mining is to segment a document into argument units before classifying them into their unit types or their stance on the topic. Existing research on argument mining takes different views on the granularity of an argument unit. The majority of existing argument mining approaches assume an argument unit to cover a sentence [9, 51, 153]. Others assume an argument unit to cover a sentence, a clause [10, 118], or multiple sentences [72].

The heterogeneity of the web in terms of the covered genres makes a sentence-level argument unit segmentation unsuitable. The reason is that genres on the web are characterized by different writing presentations and styles. Argument units in different genres cover different granularities and might not be limited to a specific syntactical unit. For example, a piece of anecdotal evidence describing personal experience might cover from a pair of sentences to multiple paragraphs. Hence, an automatic approach to segment a web document into argument units is needed. This leads to our second research question

(RQ2): *How to extract argument units in a genre-robust way?*

To answer this question, we model argument units in three corpora that cover different genres using BIO format. With BIO format, each token is tagged with one of three labels: (B)egining, (I)nside, and (O)utside of an argument unit according to the token's position with regard to the argument unit. To capture the context around each token, we analyze different semantic, syntactic, structural, and pragmatic feature types. In in-genre and cross-genre experiments, we develop and evaluate three models: a support vector machine (SVM), a linear-chain conditional random field (CRF), and a bidirectional long short-term memory (Bi-LSTM). The three models correspond to increasingly complex levels of modeling context: The SVM considers only the current token. The CRF is additionally able to consider the preceding classifications. Finally, the Bi-LSTM can exploit all words in the document. We find that the Bi-LSTM performs best in the in-genre and cross-genre experiments, which shows the importance of capturing a broad context while detecting argument units. However, the experiments show that the used models and features are insufficient for a genre-robust argument unit segmentation.

Argument mining approaches are developed based on corpora which cover a limited set of topics. However, controversial topics that users of an argument retrieval system might ask questions about are not limited to a specific list. For argument retrieval systems to serve this information needs, their underlying argument mining approaches should generalize to new topics. Argument mining approaches are developed in a supervised fashion, where an argument corpus is split into training and test sets. A supervised classifier is then trained on the training set and evaluated on the test set. The generalizability of the supervised approaches to new topics is tight to how the topics for the training and test sets are chosen.

Research on argument mining shows that existing supervised approaches learn topic-specific features [134], which hinders their generalizability to new topics. Researchers recognize the topic dependence of some argument mining tasks by choosing different topics in the training and test sets [147, 153]. While such experimental design helps to foster topic generalizability, the topics used in the training and test sets might be similar in terms of the type of arguments that target them. For example, a black market argument can be used for both "gun control" and "banning marijuana". An open research question is

(RQ3): *How to assess and foster generalizability over topic in argument mining approaches?*

To answer this research question, we create three comprehensive and authoritative sources of controversial topics and use them to assess the topic coverage of 31 argument corpora. The topic sources are created from Wikipedia, Debatepedia, and the World Economic Forum and are modeled as topic ontologies. A topic ontology is a directed acyclic graph whose nodes are topics and whose edges imply an "is part of"-relation between the topics [177]. Using the three topic ontologies, we develop manual and automatic approaches to map a given argumentative document to the topics in the ontologies that best describe the documents. We use the approaches to identify what topics the documents in the existing argument corpora cover. By analyzing the topic coverage of the argument corpora, we show that most existing argument corpora are governed by a skewed topic distribution toward a narrow set of topics.

Fostering topic generalizability in argument mining rests not only on how argument corpora are created but also on how argument mining tasks are defined. Stance classification is a task whose input is an argument and a topic and whose output is either pro or con. Providing the topic as input in terms of a short label provides little context and knowledge about the issue at stake. We introduce a variant of stance classification that we expect to foster topic generalizability: *same side stance classification*. The same side stance classification task takes a pair of arguments as input and returns whether the arguments are on the same or the opposite side. In this way, the task is no longer dependent on an input topic; hence, we expect a classifier trained on this task to generalize across topics. To analyze the generalizability of stance classification over the topic, we conduct cross-topic and in-topic experiments. In both experiments, we evaluate several classifiers and show that they generalize well across topics.

1.1.3 Identification of Argument Frames

Being effective in argumentation on a topic boils down to arguing in a way that supports the author's stance and fits the target audience [49]. Several attributes of an audience should be taken in account (e.g., emotional state, education, age). After understanding the target audience, an author should select arguments that suit the audience. Finally, the arguments should be phrased in a language that appeals to the audience.

Framing means emphasizing an aspect of a perceived reality and making the aspect more salient in a text [52]. In argumentation, a controversial topic like "nuclear energy" can be framed by emphasizing a certain theme (e.g., economy or environment). To be effective in argumentation, a speaker should choose frames that resonate with the target audience. For example, a frame like "health" might resonate more with an old audience than a young one. The topic of interest largely decides what frames can be used for. While some frames are *generic* (e.g., "economy"), other frames are specific to a given topic (*topic-specific*) [170].

Argument retrieval systems rank arguments by their relevance and quality to help users find convincing arguments [128]. Existing argument retrieval systems rank arguments without taking their frames into account, which might result in the top-ranked arguments covering only a few frames. Argument retrieval systems should equip users with an overview of the existing frames on a controversial topic. Delivering arguments with their frames helps users to navigate the arguments according to their aspects and pick those arguments that are likely to appeal to their audience. In this way, frames guide the user to navigate the space of arguments in a comprehensive way.

In Section 5.1, we approach the research question

(RQ4): *How to model and identify the frames of an argument?*

To answer this question, we propose a model that defines a frame as a set of arguments that share an aspect. Following the proposed model, we introduce an argument corpus that covers 465 topics, 1,623 frames, and 12,362 arguments. Existing framing corpora cover generic frames only [35] or few frames (only seven) [108]. In contrast, our corpus is provided with 330 generic frames and 1,293 topic-specific frames. This allows us to develop approaches that identify both types of frames and compare the effectiveness of the approaches at identifying the two frame types.

To automatically identify frames in a set of arguments, we propose an approach that consists of three steps: given a set of arguments, we first map them into a semantic space (e.g., TF-IDF) and cluster them into topics. In the second step, we remove from the argument clusters topic-specific tokens. Finally, we cluster the topic-free arguments again to produce the final clusters of frames. Using the dataset, we conduct experiments to evaluate our approach at identifying generic and topic-specific frames. To put our results in context, we compare our approach with a baseline that maps arguments into a semantic space and directly clusters arguments into frames. Our experiments clearly show the benefit of removing topic-specific features for identifying an argument's frame. In particular, our experiments show that our approach is effective at identifying generic frames, reaching a better F1-score than the baseline (0.28 versus 0.19 in the TF-IDF semantic space).

Frames allow users to locate groups of arguments that resonate with their audience. Early prototypes of argument retrieval systems deliver arguments as a list of texts, which allows for finding the most relevant arguments. Delivering arguments with frames requires an interface that presents the user with a comprehensive view of the frames along with the enclosing arguments.

To close this gap, we introduce in Section 5.2 a visual interface that maps the retrieved arguments for a query into an aspect space. We create two aspect spaces to model arguments in the visual interface. The first aspect space considers an aspect to be any controversial topic in Wikipedia's list of controversial topics.[1] In this aspect space, the association between an aspect and an argument is calculated by counting the occurrences of the aspect's text in the argument's text. The second aspect space is derived using a standard topic model (Latent Dirichlet Allocation; [25]). To visualize the retrieved arguments, we use generalized barycentric coordinates [100]. The visual interface also provides interactive functionalities that enable users to filter those arguments from the retrieved list that cover an aspect they are interested in. To illustrate the advantages of the visual interface, we present two use cases that show how the visual interface allows the user to explore or refine the retrieved arguments.

Here we summarize the main research questions approached in this thesis.

- RQ1: How to identify argumentative questions in the context of search engines?

  Argument retrieval systems have to deal with different genres on the web.

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

- RQ2: How to extract argument units in a genre-robust way?

  Argument retrieval systems have to deal with new topics which appear on a daily basis but are trained on existing topics.

- RQ3: How to assess and foster generalizability over topic in argument mining approaches?

  To be effective at persuading their audience, users select arguments from the retrieved ones based on what frames they emphasize of a controversial topic.

- RQ4: How to model and identify the frames of an argument?

1.2 Publications Record

TABLE 1.1: peer-reviewed papers by the author and their usage in the thesis.

Used in	Venue	Type	Length	Year	Publisher	Ref.
2	KI	Conference	Long	2019	Springer	[6]
	Yamen Ajjour**, Henning Wachsmuth, Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. Data Acquisition for Argument Search: The args.me corpus.* ***Best Paper Award at KI 2019					
3	SIGIR	Conference	Short	2022	ACM	[7]
	***Yamen Ajjour**, Pavel Braslavski, Alexander Bondarenko, and Benno Stein. Identifying Argumentative Questions on Controversial Topics.*					
4.2	Under Review	Conference	Long	2021	arXiv	None
	***Yamen Ajjour**, Johannes Kiesel, Benno Stein, and Martin Potthast. Topic Ontologies for Arguments.*					
4.1	ArgMining	Workshop	Long	2017	ACL	[3]
	***Yamen Ajjour**, Wei-Fan Chen, Johannes Kiesel, Henning Wachsmuth, and Benno Stein. Unit Segmentation of Argumentative Texts.*					
4.3	CEUR	Workshop	Long	2021	CEUR	[155]
	*Benno Stein, **Yamen Ajjour**, Khalid Al-khatib, Roxanne El-baff, Philipp Cimiano, Henning Wachsmuth, Same side stance classification.*					
5.1	EMNLP	Conference	Long	2019	ACL	[5]
	***Yamen Ajjour**, Milad Alshomary, Henning Wachsmuth, and Benno Stein. Modeling Frames in Argumentation.*					
5.2	EMNLP	Conference	Demo	2018	ACL	[4]
	***Yamen Ajjour**, Henning Wachsmuth, Dora Kiesel, Patrick Riehmann, Fan Fan, Giuliano Castiglia, Rosemary Adejoh, Bernd Fröhlich, and Benno Stein. Visualization of the Topic Space of Argument Search Results in args.me.*					
-	CLEF	Conference	Long	2020	CEUR	[28]
	*Alexander Bondarenko, Maik Fröbe, Meriem Beloucif, Lukas Gienapp, **Yamen Ajjour**, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. Overview of Touche Shared Task 2020: Argument Retrieval.*					
-	EACL	Conference	Long	2017	ACL	[174]
	*Henning Wachsmuth, Benno Stein, and **Yamen Ajjour**. "PageRank" for Argument Relevance.*					
-	ArgMining	Workshop	Long	2017	ACL	[173]
	*Henning Wachsmuth, Martin Potthast, Khalid Al-Khatib, **Yamen Ajjour**, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. Building an Argument Search Engine for the Web.*					

Chapter 2

Related Work

In this section, we introduce related work and background to argument retrieval systems. We start by introducing a brief overview on argumentation and how arguments are modeled. Later, we present related work to extracting arguments from natural text in Section 2.3. We report on related research to argument retrieval in Section 2.4. Finally, we introduce the concept of framing and computational approaches for identifying frames in natural language.

2.1 Argumentation

Argumentation is the communicative activity of forming arguments and exchanging them to resolve a disagreement of opinions [168]. We can distinguish between two types of argumentation: dialogue and monologue. A monologue is a speech given by one person to an audience without an immediate contention. In a dialogue, two or more sides interactively participate in a discussion to support their stance on the topic. A stance on a topic like "We should legalize marijuana" can be either pro the topic (meaning, "yes, let us legalize it") or con the topic ("no, let us make it illegal").

There are different goals for an argumentation which can be deciding for the best course of action on a certain matter (deliberation) or trying to ensure that the audience will respond to the speaker's message in the desired way (persuasion). Regardless of the type of argumentation or its goals, a clear need for people when involved in argumentation is constructing and finding good arguments that reach the intended effects on an audience (e.g., persuading an audience with a certain stance).

Analyzing and evaluating arguments have been studied from different angles: rhetoric, dialectic, and logic. Logic studies the internal parts of an argument in terms of a premise and a conclusion, which are propositions that can be true or false. The focus in logic is on how to formally show that the conclusion of an argument follows from its premise. Rhetoric studies how a speaker can deliver a persuasive speech to an audience. In his theory on Rhetoric (the art of persuasion), Aristotle [14] introduced three modes which a speaker should consider while delivering a speech: logos, pathos, and ethos. While ethos is concerned with establishing the credibility of the speaker, logos is concerned with employing the rules of sound reasoning. Pathos, on the other hand, deals with forming arguments in a manner that seeks to evoke an emotional response in the audience.

Dialectic is a branch of logic that analyzes arguments given in a discourse by examining criticism of them. The pragma-dialectics paradigm introduces normative rules for how constructive argumentation should be conducted in order to achieve consensus [168]. Pragma-dialectics views persuasion as strategical choices in a discourse that are related to three aspects: topical potential, presentational devices, and audience demand [49]. According to this theory, effective argumentation boils down to choosing from the topical potential, selecting presentational devices, and meeting the audience demand. Choosing from the topical potential can be understood as formulating one's stance in a way that is likely to be accepted by the audience. Meeting the audience demand means considering the beliefs and values of the audience while creating arguments. Presentational devices that help to craft arguments are rhetorical figures [11, 91] and framing. In Section 2.5, we introduce related work that includes a definition of framing and related work to framing in natural language processing.

Modeling arguments computationally is an active research area in artificial intelligence, which is called computational argumentation. Typical tasks in computational argumentation are extracting arguments from natural text and analyzing their quality. Argumentation in natural language is studied in different genres. A genre is a set of texts that have a common linguistic function (e.g., editorials or argumentative essays). Depending on the genre and application, an argument model is used to capture the concept of argument, its units, and context. An argument is defined as a set of argument units, where an argument unit is a span of text that plays an argumentative role (e.g., conclusion). In the following, we review several argumentative genres and how arguments are modeled in them.

2.2 Argument Models

In this subsection, we introduce several argumentative genres and examples of argument models used in each genre.

Argumentative Essays Argumentative essays are usually written by students in an educational context to improve their writing and critical thinking. Researchers usually develop approaches to score different quality dimensions of argumentative essays (e.g., organization [126]). Modeling arguments in argumentative essays has been approached by Stab and Gurevych [148]. In this work, Stab and Gurevych [148] let annotators label a corpus of 402 argumentative essays with four argument units: premise, claim, major claim, or other. The boundary of an argument unit is defined to be any span of text within a sentence. A premise is defined as evidence that underpins the validity of a claim. A claim is a statement that is either true or false and should not be accepted without additional support. A major claim is a statement that carries the author's stance toward the topic of the essay. To capture argument structure, the notion of argumentative relation is introduced, which indicates a support or attack relation between two argument units (e.g., premise and conclusion).

Wikipedia Wikipedia is the largest online encyclopedia that is open to contributions from anybody with internet access. Wikipedia articles cover a wide spectrum of knowledge, including concepts that are controversial. By focusing on controversial topics, Aharoni et al. [2] annotated 586 articles on 33 topics in Wikipedia with an argument model that consists of two unit types: claim and evidence. Evidence is further categorized into anecdote, statistics, or testimony. Whereas statistics is evidence in terms of quantitative or empirical data, a testimony is evidence in terms of a proposition that is made by an expert or an authority. An anecdote is an example or description of a specific event or an instance.

A key difference between this model and other argument models is the integration of the topic of an argument in the model. In this way, a premise or a conclusion is always tight to a certain controversial topic that defines the context of the argument. Argument mining approaches that build on this corpus develop computational approaches that take the topic as input [92].

Legal Texts Law is a significant source of argumentation, given the need to support court decisions, legal cases, or accusations with arguments. Early work on modeling arguments in natural text was conducted on legal texts that originate from the European Court of Human Rights (ECHR, [115]). In this corpus, legal texts represent case law, including information about a case, arguments from the contending parties (defendant and plaintiff), as well as the decision of the court. Arguments in this line of research are modeled as a conclusion that is supported by multiple premises or other arguments. On the language level, argument units are annotated on the sentence level.

Scientific Papers Argumentation is an essential element of any scientific publication. Teufel [159] coined the term argumentative zones, which are roles that passages play as a part of a paper's contribution [159]. In this work, argumentative zones in scientific papers were classified into background (generally accepted scientific background), conclusion, aim (research goal), and own (description of author's work). Lauscher et al. [87] modeled argument units on the sub-sentence level into background claim, own claim, and data. Three different relation types were distinguished between two argument units: support, attack, and semantically the same.

Editorials Editorials are opinionated news articles that seek to convince an audience with a certain view. Compared to argumentative essays, editorials are written by trained writers to advocate a certain stance on a controversial topic. To capture arguments in editorials, Al-Khatib et al. [10] annotated argument units in 300 editorials with six types: common ground, assumption, testimony, statistics, anecdote, and others. While common ground is knowledge about a topic that is accepted by everybody, assumptions are statements that carry the stance of the author. Testimony, statistics, and anecdote are types of evidence that are similar to those introduced by Aharoni et al. [2]. Other is used when none of the argument unit types match.

Debate Portals Debate portals are websites where people debate a controversial topic or outline the main arguments on it. Debate portals can be divided into dialogical, which allows two opponents to debate a certain topic, and monological, where pro and con arguments are listed. Debate portals constitute a primary source for studying argumentation, given the high quality of arguments and ease of acquiring them. Early work on debate portals was conducted by Cabrio and Villata [33], who developed an approach to detect whether an argument in a debate supports or attacks the topic. In this work, an argument is modeled as a span of text that holds support or attack relations to other arguments on the topic.

2.3 Argument Mining

Argument mining aims at extracting arguments from natural language. An argument mining pipeline first identifies spans of text that are argumentative (argument unit segmentation). Later, these text spans are classified into multiple labels: premise or conclusion (argument unit classification). Argument relation identification aims at detecting supporting and attacking relations between two argument units. A closely related task to argument relation identification is stance classification, where the stance of an argument unit or an argument is detected with regard to a given topic. In the following, we report related work on each of these tasks.

2.3.1 Argument Unit Segmentation

Unit segmentation is a classical segmentation task, which is related to discourse segmentation [15, 64, 121] as for rhetorical structure theory [98]. Both discourse and argument units are used as building blocks, which are then hierarchically connected to represent the structure of the text. However, argument units are closer to classical logic, with each unit representing a proposition within the author's argumentation.

Much existing work on argument mining skips the segmentation, assuming segments to be given. Such research mainly discusses the detection of sentences that contain argument units [104, 115, 139, 159], the classification of the given segments into argumentative and non-argumentative classes [149], or the classification of relations between given units [120, 122, 149].

A few publications address problems closely related to unit segmentation. Madnani et al. [97] identified non-argumentative segments, but they did not segment the argumentative parts. Levy et al. [92], on the other hand, tried to detect segments that are argumentatively related to specific topics. However, they did not segment the whole text.

A unit segmentation algorithm has been applied already by Al-Khatib et al. [10] in the creation of the Editorials corpus analyzed in Section 4.1. The authors developed a rule-based algorithm to automatically pre-segment the corpus texts before the manual annotation. The algorithm was tuned to rather split segments in cases of doubt. During the annotation, annotators were then asked to correct the segmentation by merging incorrectly split segments. The authors argue that—even with a simple algorithm—this approach simplifies the annotation process and makes evaluating inter-annotator agreement more intuitive.

In the few publications that fully address unit segmentation, a detailed analysis of features and models is missing. Previous work employs rule-based identification [124], feature-based classification [90], conditional random fields [141, 154], and deep neural networks [50]. Especially early approaches by Stab [154] and Eger et al. [50] relied on sophisticated structural, syntactical, and lexical features. Eger et al. [50] even report that they beat the human agreement in unit segmentation on the one corpus they consider. Still, the paper does not clarify which linguistic cues are most helpful to reach this performance. In Section 4.1, we also employ a deep neural network based on Bi-LSTM, but we perform a detailed comparison of models and feature sets.

Most existing work trains and tests unit segmentation algorithms on one single corpus. A frequent choice is one of the two versions of the Argument Annotated Essay Corpus [148, 154], which is studied by Persing and Ng [124], Eger et al. [50], Stab [154] himself, and also by us. However, for a unit segmentation algorithm to be used for web documents, it has to work robustly also for texts from other genres. Section 4.1, therefore, extends the discussion of unit segmentation in this direction.

Among the few works that study unit segmentation in a cross-genre setting is [165]. In this work, the authors modeled an argument as a span of text that is either pro, con, or neutral to a topic. In this way, argument unit segmentation is performed jointly with stance classification. A corpus was created by retrieving the top 500 documents from the Common Crawl Index[1] on eight controversial topics. Then, several sequence-to-sequence classifiers were developed to perform the task both in a cross-topic and in-topic settings. The results of the experiments show very close results between the in-topic and cross-topic settings, reaching a drop of 0.06-0.07 F1-score.

Performing unit segmentation with unit classification was studied in a multi-task setup [143]. Similar to our approach, sequence-to-sequence models were trained to detect whether a token is the (B)eginnig, (I)nside, or (O)utside an argument unit. In contrast to our approach, BIO encoding was done depending on the argument unit type. For example, the beginning token of a conclusion (respectively premise) was encoded as Conclusion-B (respectively Premise-B). Our encoding scheme, however, encodes the beginning of a token as Arg-B regardless of its type. The used sequence-to-sequence model is a combination of Bi-LSTM and CRF [132]. This model was evaluated in two settings: a single-task and a multi-task setup. In the single-task setup, the model was tested and evaluated on one of five corpora that cover the genres: Wikipedia, web, news comments, argumentative essays, and various. In the multi-task setup, the model detected argument units in the five corpora jointly by having a shared Bi-LSTM neural network but a separate CRF layer on top for each different corpus. The experiments show that the model performed better in the multi-task setup, indicating that training on multiple genres benefits argument unit segmentation.

2.3.2 Argument Unit Classification

Argument unit classification is the task of classifying an argument unit into one of the multiple types defined by the adopted argument model (e.g., conclusion or premise). The argument unit types are decided by the argument model, and hence, the task is dependent on how arguments are modeled. Researchers train supervised classifiers to categorize an argument unit into one of the pre-defined argument unit types. Early approaches for argument unit classification used supervised feature-based classifiers to categorize argument units in specific genres. Typical genres which are studied for argument unit classification are legal domain [105], essays [148]¸ discussion forums [117], editorials [10], scientific papers [87], and debate portals [73].

Feature sets used for the classifiers involve syntactic and lexical linguistic clues (e.g., part-of-speech tags or sentence length) as well as context information of the argument unit. Context information is captured by features that are extracted for the preceding and following argument units [115, 149]. The classification performance is measured in terms of macro-averaged F1-score over the argument unit types. Examples of classification effectiveness is 0.72 on argumentative essays [149] and 0.64 on debate portals [73].

[1] `http://index.commoncrawl.org/CC-MAIN-2016-07`

Cross-genre argument unit classification has been first approached by Daxenberger et al. [44] and Habernal and Gurevych [71]. Habernal and Gurevych [69] used support vector machines (SVM; [41]) to classify argument units in documents from different genres (discussion forums, news, and blogs) into the unit types of the Toulmin Model. The unit types include evidence, claim, refutation, rebuttal, and backing. The classification performance in the cross-genre setting is an F1-score of 0.21. This low classification performance can be justified by the difficulty of applying an elaborate annotation scheme, such as the Toulmin model, on the web.

A more simplistic approach was introduced by Daxenberger et al. [44], who classified sentences into claim or non-claim. Similar to our experimental setting in Section 4.1, the experiments involved six argument corpora that represented six genres (e.g., argumentative essays and news). In the cross-genre setting, classifiers were trained on one of the argument corpora and tested on another corpus. The classifiers used in the experiments were logistic regression and a convolutional neural network. Comprehensive feature sets were used to represent an input sentence which included lexical, syntactic, semantic, structure, and discourse features. The experiments show that simple lexical features in terms of token n-grams are the best indicators of claims in and across genre. The F1-score of the logistic regression classifier in the cross-genre setting is subpar to that in the in-genre setting (a drop of 0.14 points in terms of claim F1-score).

The emergence of transformers allowed for a more effective classification of argument unit types. Ein-Dor et al. [51] introduced an approach to detect evidence in a corpus that first filters sentences for a given topic and then ranks sentences using BERT [46]. To filter sentences on a topic, queries are formulated using the topic and lists of sentiment and evidence-related words. Then, a three-step approach is iteratively run to detect evidence gradually: First, run a classifier on the corpus. Then, let annotators label the classified sentences in the corpus. Finally, retrain the classifier on the newly labeled sentence set. The approach was applied to a news corpus and resulted in 198,457 sentences, among which 33.5% are labeled as evidence. The corpus was used in several experiments to train a BERT classifier with the goal of detecting evidence sentences in the same corpus and in different corpora. The experiments show that BERT can rank the top 20 candidates with a precision of 0.95 and 0.85 in the in-genre and cross-genre settings respectively. These results show promising applications in argument retrieval systems.

2.3.3 Argument Relation Identification

Argument relation identification is the task of identifying whether an argument unit supports or attacks another argument unit. Relations between argument units in a text make up a tree whose nodes are the argument units and whose edges are the argument relations between them. Palau and Moens [115] developed context-free grammar to parse argument structures in legal documents. While such an approach can work for specific genres, developing context-free grammar that work across genre is rather hard. Stab and Gurevych [149] trained an SVM on argumentative essays to identify whether an argument unit supports another argument unit or not. The best distinguishing features are lexical Boolean features that include, for example, word pairs in the argument unit pair and modal verbs.

A similar observation was made by Lawrence and Reed [89], who used several features that include topic similarity and discourse markers to train classifiers for argument relation identification. Topic similarity between an argument unit pair was captured using the length of the shortest path in WordNet[2] between all word pairs in the argument unit pairs. According to experiments on AIFdb [22], which covers several genres, the topic relatedness of an argument unit pair is a stronger indicator of their support or attack relation than discourse makers.

Apart from topic knowledge, approaches that jointly perform other argument mining tasks show higher effectiveness at the task of argument relation identification [164]. Peldszus and Stede [122] used a Minimum Spanning Tree to identify argument relations in short argumentative essays. The approach starts by training classifiers to predict the types of argument units and their relations. The Minimum Spanning Tree was then applied on a graph whose nodes are the argument units and whose edges are the support/attack relations. The output of the classifiers on the separate tasks was then combined to derive weights for the edges between the pre-segmented argument units in a piece of text. The approach shows that jointly performing the tasks outperforms running them separately.

2.3.4 Stance Classification

Stance classification deals with identifying whether an argument unit expresses an attitude in favor or against a given topic. Stance classification has been approached in several genres, including debate portals [146], social media [107], argumentative essays [54], Wikipedia [19], and news [55]. The topic input to the task can be a statement or a short phrase (e.g., a product or a topic). Some researchers modify the label set for stance detection by adding further labels. For example, in fake news detection [74], a label is added to describe texts as irrelevant for a given target. The "neutral" label is usually added in genres where the texts are not always argumentative (e.g., in news articles [55]).

A major challenge in stance classification is that a text might not mention an explicit stance towards a topic of interest but only expresses an implicit stance [176]. For example, for a question like "Should vaccines be mandatory?", an answer might be "Medical consent is protected by human rights." Even though this answer does not explicitly mention a stance toward vaccines, it can easily be inferred

[2] http://wordnet.princeton.edu/

that the answer is against mandatory vaccines (implicitly mentioned by “medical consent”). Wojatzki and Zesch [176] studied implicit stance patterns in tweets towards the topic “atheism” by analyzing stance labels towards explicitly mentioned topics (e.g., “Islam” or “Christianity”). Utilizing the explicit stances as features to detect the stance toward the topic of interest outperformed the state-of-the-art on this task. Bar-Haim et al. [20] annotated 3,000 pairs of topics as consistent or contrastive, depending on whether people have the same stance towards the target pairs (e.g., e-mobility and electric cars) or the opposite stance (e.g., fuel cars and electric cars).

2.3.5 Role of Topics in Argument Mining

Extracting and analyzing arguments automatically from natural text benefits from defining and incorporating relevant topic knowledge. Levy et al. [92] and Rinott et al. [136] consider arguments to be topic-dependent and study their detection in the context of a random selection of up to 58 topics from `idebate.org`. This work raises the question of why topic-dependence has not been addressed more urgently until now.

Integrating topics in supervised classifiers helps identifying arguments and their stance more effectively [151]. The approach of Stab and colleagues is a modified version of Bi-LSTM [63], which incorporates the topic while jointly detecting (1) whether a sentence is an argument and (2) its stance on the topic. The designed neural network outperforms Bi-LSTM without topic integration in both tasks; the approach gives further evidence for the topic-dependence of argument mining and stance classification. Whether model transfer between more closely related topics works better is unknown. In Section 4.2, we introduce three topic ontologies that are tailored to arguments and formally capture relatedness relations between topics.

Building on the work of Stab et al. [151], Fromm et al. [57] developed a neural network that embeds the words in a topic using word embeddings and knowledge graph embeddings. To derive knowledge graph embeddings, the words of the topic are matched with a knowledge graph (DBpedia). Both the sentence and the topic are forwarded to two bidirectional recurrent neural networks and then aggregated and forwarded to a multi-layer perceptron. The topic dependence of stance classification was recognized by Reuver et al. [134], who showed that cross-topic stance classification with BERT [46] produces mixed results depending on the topics. In Section 4.3, we introduce a new formulation of stance classification (same side side stance classification) that takes a step toward making the task less dependent on the topic as input.

Incorporating topic knowledge in supervised approaches for argument relation identification is an active research direction. Paul et al. [119] used a knowledge graph (ConceptNet[3]) to extract entities in an argument unit pair and the graph paths between all entity pairs. Knowledge graphs include factual knowledge in terms of edges that represent two entities. An example can be (marijuana, is a type of, drug). The entities and paths are embedded in a neural network together with the text of the argument units. The proposed neural architecture outperforms several strong baselines, scoring an F1-score of 0.64 on debate portals and 0.60 on argumentative essays.

Robustness to unseen topics was first approached by Cabrio and Villata [33], who used a textual entailment system to identify argument relations in debate portals. The textual entailment system classifies an argument to either entail (i.e., support) or contradict (i.e., attack) another argument on the topic. The approach achieves an accuracy of 0.69. This approach for argument relation identification is close to same side stance classification, which we introduce in Section 4.3. However, we conduct in-topic and cross-topic experiments to evaluate the generalizability of approaches across topics. While the experiments conducted by Cabrio and Villata [33] evaluate the system on unseen topics, the used dataset was rather small, covering 200 arguments and 19 topics. A more elaborate and comprehensive topic sampling while designing experiments is likely to reveal more insights into how difficult the task is. In Section 4.2, we introduce topic resources that are suited for argumentation and that are created by domain experts.

Topic knowledge is also utilized to generate arguments. Bilu et al. [23] introduced 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. [20] identified 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. Perhaps only abstract argumentation can be conceived as topic independent, since it studies the structure and relations among arguments more than their language.

Topic Ontologies In information science, an ontology is defined as "an explicit specification of a conceptualization" [66]. Topic ontologies are a specific type of ontologies that specify topics as nodes of a directed acyclic graph. An edge in the graph then implies an "is part of"-relation between the topics [177]. 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 the ACM

[3] `https://conceptnet.io/`

Computing Classification System for computer science or DMOZ for web pages.[4] The only topic ontology directly linked to arguments is that of Debatepedia.

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) [144]. Researchers usually train supervised classifiers for each class in the hierarchy [156].

2.4 Argument Retrieval

Argument retrieval is a research area centered around the idea of retrieval systems that retrieve pro and con arguments for a given query. Queries related to argument retrieval are those that seek opinions on a topic. In information retrieval, queries related to controversial topics and forming opinions have been approached from several angles, which we review first. Research on argument retrieval started by proposing prototypes that envision architectures for argument retrieval. The prototypes take pragmatically different views, which we systematically compare in Subsection 2.4.2. A central component in an argument retrieval system is a ranking model that sorts arguments for a given query according to a specific quality criterion. We report in Subsection 2.4.3 on approaches that cover ranking arguments with regard to their quality and topical relevance.

2.4.1 Argumentative Questions

Argumentative questions look for arguments or opinions on a given topic. In this thesis, we introduce an approach to identifying argumentative questions in the query logs of search engines. Related work to argumentative questions covers analyzing search engine bias on controversial topics, analyzing queries on controversial topics, and identifying opinion questions in community question answering.

Bias in search engines Starting from a pre-defined list of controversial topics, Gezici et al. [60] observed a tendency by major search engines to rank liberal content higher than conservative content. Kulshrestha et al. [84] evaluated the contribution of search system components such as source documents and ranking algorithms to political bias in search results. Yom-Tov et al. [180] observed that most users are more likely to read opinions that match their views, and that diversification of the results can only be successful if the documents with the opposite view are lexically similar to the user's queries. Azzopardi [16] surveyed different sources of cognitive bias in search on socio-political topics. Our work sheds light

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

on the count and characteristics of argumentative questions in query logs, whose askers might be more prone to bias than those of factual questions.

Query analysis on controversial topics Gyllstrom and Moens [67] identified controversial topics in web search queries by checking whether Google auto-complete suggests positive and negative words for a given concept. Weber et al. [175] filtered queries on controversial topics and labeled them with "left" or "right" by checking whether the clicked URL for a query is a left or right political blog. Chelaru et al. [36] extracted queries from AOL query logs using templates and labeled them as positive, negative, or objective. Topics that occur in both positive and negative queries more than a specific threshold were considered to be controversial. While this work is closest in spirit to our approach for identifying argumentative questions in Section 3, our work focuses on question-like queries which are less ambiguous with regard to their intent than short queries. Cambazoglu et al. [34] annotated a sample of 1,000 web search questions with a taxonomy of 16 question types, which include opinion and reason questions. The study shows that opinion and reason questions amount to about 1% of web search questions. The low magnitude of opinion questions in search query streams is a challenge for such holistic taxonomies and hence for developing answering systems for them. In Section 3, we concentrate on argumentative web search questions by sampling questions on controversial topics and classifying them into factual, argumentative, and method questions.

Opinion questions in CQA/QA Researchers studied subjective and opinion questions mainly in the context of community question answering (CQA). The motivation for this classification was to fact-check answers to factual questions [101], suggest similar questions with the same intent [37, 94], or automatically answer opinion questions [18, 83]. The latter approaches classify the polarity of an opinion question into negative or positive and return answers with the same polarity for the question. Since this might lead to undesired bias by emphasizing the view of the question, Moghaddam and Ester [106] proposed a mechanism to retrieve both negative and positive answers. Even though opinion questions seem to be close to argumentative questions, argumentative questions differ conceptually by targeting controversial topics, which are characterized by disagreement and require reasoned arguments as answers.

Argumentative questions and controversial topics are a source of bias for search engines. Argument retrieval systems provide the needed technology to respond appropriately to argumentative questions. In the following, we compare three existing argument retrieval systems with regard to the choice of their argument sources and the consequences thereof on the design of each of the argument retrieval systems.

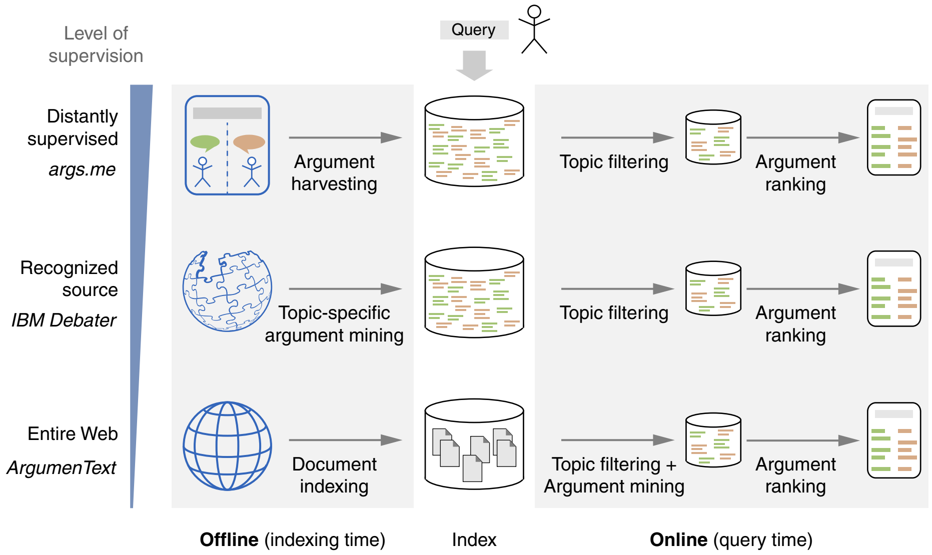


FIGURE 2.1: Comparison of three general argument acquisition paradigms: args.me and IBM Debater index arguments offline, relying on distantly supervised harvesting and on mining from recognized sources respectively. ArgumenText indexes documents and mines online at query time. The level of supervision reflects the effort humans spend to create arguments from a source, which in turn implies notable differences regarding index sizes, topic bias, and noise in the data.

2.4.2 Argument Acquisition

Existing argument retrieval prototypes [93, 151, 173] follow paradigmatically different approaches to argument acquisition: see Figure 2.1 for a comparison. The choice of argument sources and mining methods is usually tightly coupled and constitutes a decisive step in designing an argument retrieval system. The smaller the ratio of explicit arguments to other text in the sources, the more effort needs to be invested to mine high-quality arguments.

ArgumenText (Figure 2.1 bottom) follows web search engines in indexing entire web documents. Using a classifier trained on documents from multiple genres, ArgumenText then mines and ranks arguments from topically relevant documents at query time [152]. The advantages of this approach are recall maximization ("everything" is in the index) and the possibility to decide whether a text span is argumentative on a per-query basis. A disadvantage may arise from the aforementioned as of yet unsolved problem of cross-genre robustness [44].

The approach of IBM Debater (Figure 2.1 center) is to mine conclusions and premises of arguments from recognized sources (such as Wikipedia and high-reputation news portals) with classifiers trained for specific topics [92, 93, 136]. The arguments are indexed offline (i.e., unlike ArgumenText, the retrieval unit is an argument, not a document)—the complete documents may still be stored in an additional storage. Argument retrieval then boils down to topic filtering and ranking. While the source selection benefits argument quality, recall depends on the effort invested into the training of the classifiers (i.e., human labeling is involved to guarantee the effectiveness of the topic-specific classifiers).

Finally, the approach of args.me is shown at the top of Figure 2.1. Arguments from debate portals are indexed offline, similar to IBM Debater. However, instead of classifier-based mining, arguments in args.me are harvested using distant supervision, exploiting the explicit debate structure provided by humans (including argument boundaries, pro and con stance, and meta data). This does not only benefit the retrieval precision but also renders our approach agnostic to topics. A shortcoming of this approach is that it needs to decide what an argument is at indexing time, independent of a query. To some extent, this restriction can be overcome in the future through more elaborated topic filtering and ranking algorithms. Besides, the gain of precision comes at the expense of recall as the number of sources qualifying for distantly-supervised argument harvesting is limited.

2.4.3 Argument Ranking

Ranking arguments should take their quality in account. Aristotle [14] introduced in his theory on the art of persuasion three modes that a speaker should consider while delivering a speech: logos, pathos, and ethos. While ethos is concerned with establishing the credibility of the speaker, logos is concerned with constructing or delivering a logical argument. Pathos, on the other hand, deals with appealing to the emotions of the target audience.

According to the theory of pragma-dialectics, the quality of an argument emanates from framing it in a way that fits the audience. The values of the audience and the sequence in which arguments are delivered play a major role in their effectiveness on the audience. In Section 5.1, we introduce an approach to identify the frames of an argument, which enables users to select and arrange arguments based on the audience they are targeting.

Logic evaluates an argument with regard to its constituting argument units and their relations. In informal logic, an argument is considered good if it satisfies three conditions: acceptability, relevance, and sufficiency, and in this case, it is called cogent. An argument is considered acceptable if its premises are worthy of being considered true by a given audience [24]. The premises of an argument are considered relevant to its conclusion if the acceptance of the premises has some bearing on the truth of the conclusion [43]. Notice that a global notion of argument relevance exists, which represents its contribution to resolving the disagreement tackled in an argumentation. Sufficiency assesses whether the premises of the arguments provide enough amount of grounds to the conclusion [24].

Ranking arguments raises the question of what quality dimensions should be considered while developing and evaluating retrieval models for arguments.

Wachsmuth et al. [172] systematically categorized three aspects of argument quality into a taxonomy that covers logic, rhetoric, and dialectic. Each of the aspects is subdivided into several dimensions (e.g., logic is subdivided into local relevance, local acceptability, and local sufficiency). Using the dimensions, seven annotators labeled 320 arguments with regard to 15 quality dimensions with a three-point scale. The inter-annotator agreement on the 15 dimensions ranged between 0.26 and 0.51 Krippendorfs' α, showing the subjectivity of assessing argument's quality. The study shows that humans agree more on how reasonable an argument is than how well it appeals emotionally to an audience (0.5 for reasonableness versus 0.26 for emotional appeal). Wachsmuth and Werner [171] extended this study by developing regression models to rank the labeled arguments using features such as the usage of discourse markers. The regression models achieve an acceptable mean absolute error ranging from 0.22 to 0.44 for the different quality dimensions. According to the experiments, the best features to predict argument quality are subjectiveness as modeled by the usage of personal pronouns, sentiment, as well as the length of the argument text.

To assess global argument relevance, Wachsmuth et al. [174] created a graph of arguments by connecting two arguments when one uses the other's conclusion as a premise. Later, they exploited this structure to rank the arguments in the graph using PageRank [114]. This method is shown to outperform several baselines that only consider the content of the argument and its local structure (conclusion and premises).

A departure from the theory on argument quality was first taken by Habernal and Gurevych [70], who let crowd workers choose which argument in an argument pair is more convincing. Labeling arguments in pairs provides a reference for comparison as well as more context about the topic, making the ranking arguments easier for annotators and supervised approaches. The annotation study covered 16,081 argument pairs and 16 topics. Additionally, the crowd workers provided a reason for why the chosen argument is more convincing. The authors derive a ranking for all arguments on a topic by constructing a directed argument graph and applying PageRank to it. The nodes of the constructed argument graphs are arguments, while an edge implies that the source argument is more convincing than the target argument. Supervised classifiers such as support vector machines and transformers achieve an accuracy of 0.78 and 0.83 on this task [160] respectively. The high classification performance shows the feasibility of ranking arguments according to their persuasion using machine learning approaches.

From an information retrieval perspective, retrieval models are the standard components in a retrieval system for ranking documents. A retrieval model is a function that takes a document and a query and returns a score representing how relevant the document is to the query [42]. Retrieval models rank documents according to their topical relevance to the query and their user relevance [42].

Whereas the topical relevance of a document is whether it is about the same topic of the query, user relevance is more concerned with the document's contribution to solving the problem faced by the user [178]. The first studies on argument retrieval used standard retrieval models to rank arguments. The first prototype of an argument retrieval system used BM25F to rank arguments, which allows giving different weights to the conclusion, premise, or context of an argument.

Potthast et al. [128] evaluated four retrieval models at ranking arguments with regard to their relevance to a given query as well as argument quality. For argument quality, the authors adopted the argument quality aspects surveyed by Wachsmuth et al. [172]: logic, rhetoric, and dialectic. One of the main findings is that DirichletLM [181] and DPH [12] are on par, and each is better at ranking arguments than BM25 [137] and TF-IDF [80]. Gienapp et al. [61] extended this work by proposing a pairwise strategy that reduces the costs of crowdsourcing argument retrieval annotations by 93% (i.e., annotating only a small subset of argument pairs). Bondarenko et al. [29] introduced Touché which is a shared task that covers two tasks: retrieving arguments for argumentative questions and retrieving argumentative documents (i.e., web documents that include arguments) for comparative questions. The approaches submitted to the shared task integrate techniques for assessing argument quality and query expansion.

Dumani et al. [48] introduced a probabilistic framework that operates on semantically similar conclusions and premises. The framework utilizes support/attack relations between clusters of premises and conclusions and between clusters of conclusions and a query. The framework is found to outperform BM25 in ranking arguments. Later, Dumani and Schenkel [47] also proposed an extension of the framework to include the quality of a premise as a probability by using the fraction of premises that are worse with regard to the three quality dimensions of logic, rhetoric, and dialectic. Using a pairwise quality estimator trained on the Dagstuhl-15512 ArgQuality Corpus [172], their probabilistic framework with the argument quality component outperformed the one without it on the Touché Task 1 data [27].

An important aspect of argument ranking approaches is guaranteeing that the retrieved arguments are not biased toward a certain stance on the topic. Cherumanal et al. [38] introduced three fairness metrics that evaluate whether a certain stance is more dominant than the other in a ranking of arguments. The more arguments are found to be supporting a specific stance in the top positions of the retrieved arguments, the lower the reported fairness by the metrics. The metrics are used to evaluate the fairness of the rankings by all systems submitted to Touché 2020 [27]. The correlation between the NDCG of the submitted systems and the three fairness metrics is negative, showing that the output of argument retrieval systems can be biased toward a particular stance.

Use cases for argument retrieval include writing and debating support. In comparison to user queries in conventional search that can often be satisfied by one or a few retrieved documents, these use cases require a broader consideration of the retrieved arguments. Hence, users of an argument retrieval system will often investigate both stances and multiple frames on a given topic. While several studies tackled the task of ranking arguments according to their quality [69, 174], how to aggregate arguments into frames is largely unstudied. In Chapter 5, we introduce an approach to identify the frames of an argument and present a visual interface that presents arguments along the aspects they cover of a controversial topic. In the following, we introduce the concept of framing and related work to framing in natural language processing.

2.5 Framing

Entman [52] was the first to introduce a formal definition of framing as a way to select and make specific aspects of a topic salient. Subsequent research on framing is concentrated on the effect of using frames in news articles on a specific audience. One of the open questions is whether frames are topic-specific, generic, or both. Vreese [170] studied framing in news articles and considered frames to be either topic-specific or generic. Johnson et al. [79] and Card et al. [35], on the other hand, defined frames to be independent of the topic and investigated their usage across different topics.

Recently, framing caught some attention in the NLP community. Different computational models have been developed for modeling frames in natural language text. Recasens et al. [131] analyzed frames in general by identifying one-sided tokens in Wikipedia articles. Tsur et al. [166] used topic models on statements released by congress members of the two major parties in the US, Republicans, and Democrats. The learned topics were then aggregated into clusters, such as health and economy, and interpreted as being generic frames. On this basis, the authors studied the frequency of the frames in the released statements for the two parties as well as their distribution over time. Related work was conducted by Menini et al. [99] to model frames in political manifestos released by the parties (texts declaring a stance) as clusters of key phrases. The developed method was shown to outperform standard topic models in capturing frames.

Card et al. [35] annotated around 20,037 news articles on three topics (same-sex marriage, immigration, and smoking), along with a list of 15 generic frames. While the annotations had to cover continuous text spans, their granularity was left unspecified. The inter-annotator agreement on frames for the different frames ranged between 0.08 and 0.23 in terms of Krippendorff's α. By comparison, our dataset covers both generic and topic-specific frames and is annotated on the argument level. Naderi and Hirst [109] extended the work of Card et al. [35] by training a neural network to classify the frames in the constructed corpus. The authors modeled frames on the sentence level and reached an accuracy of 53.7% in multi-class classification and 89.3% for one-against-others classification. Using the same corpus, Field et al. [56] created a lexicon for each one of the 15 frames and analyzed which frames are used mainly to talk about the United States in Russian news.

The relation between arguments and frames was introduced briefly in some works [31, 58]. Still, recent research on computational argumentation largely ignores frames, and a model for aggregating arguments into frames is still missing. Naderi [108] considered a frame to be an argument and classified sentences in parliamentary speeches into one of seven frames. Reimers et al. [133] created a dataset of argument pairs that are labeled according to their similarity. Based on the dataset, they introduced the task of argument clustering, which aims at classifying an argument pair with the same topic into similar or dissimilar. The main difference to our work in Chapter 5 is that no explicit aspect is assigned to the arguments during annotation.

Chapter 3

Identifying and Analyzing Argumentative Questions

Information needs related to forming an opinion on a controversial topic constitute a challenge for search engines. An example of such a question is “why should marijuana be legalized?” Search engines are less effective in answering such non-factual questions compared to factual ones [34]. Moreover, search results for queries related to controversial topics tend to be a source of bias [60].

Retrieving arguments to a controversial topic promotes transparency since arguments not only support a position on a controversial topic but also include the justification for this position. In this section, we tackle the task of identifying questions that look for arguments—argumentative questions—in the query log of a search engine using a two-step scheme. Our scheme simplifies the task by first detecting whether the context of a question is controversial or not, and, if the context is controversial, then classifying the question as one of factual, method, or argumentative. Using the annotation scheme, we perform a crowdsourcing annotation of a sample of the Yandex query log from 2012. This results in a dataset of 39,340 questions about 19 controversial topics.[1] We analyze the questions in the dataset with regard to their form in order to gain insights into what the characteristics of argumentative questions are. To operationalize our method, we build classifiers to automatically identify argumentative questions along with factual and method ones.

3.1 Dataset Construction

Given that no available question datasets exist on controversial topics, we conducted an annotation task to create one starting from 2 billion archived Yandex queries in Russian from 2012.

Following Völske et al. [169], we first extracted queries from the Yandex log starting with a question word that resulted in 1.5 billion questions. To find ques-

[1] The anonymity of the questions’ askers is preserved by sampling only frequent questions from the logs. An exemplary data sample can be found here:
https://files.webis.de/data-in-production/data-research/arguana/webis-arg-questions/dataset.csv

TABLE 3.1: Controversial topics used in the study.

Debate portals	Russian news
Abortion	2011−2013 Russian protests
Death penalty	Alexei Navalny
Euthanasia	Anatoliy Serdyukov
Evolution	European debt crisis
Gay marriage	Floods in Krymsk
God exists	Magnitsky Act
In vitro fertilization	Nord Stream
Legalize marijuana	Presidential elections
	Putin
	Pussy Riot trial
	Yukos

TABLE 3.2: The absolute and relative count of the questions per label in the dataset.

Label	Absolute	Relative	Label	Absolute	Relative
Topic abountness			Question types		
On topic	40,689	73%	Factual	25,332	64%
Not on topic	11,665	24%	Argument.	10,982	28%
Ill-formed	1,477	3%	Method	3,026	8%

tions asking about controversial topics, we created a list of such topics (cf. Table 3.1) by: (1) Selecting eight debate topics from the args.me corpus [6] with the highest number of arguments. (2) To cover local issues, we also selected 11 debated topics from the list of the most important events in 2012 according to the Russian RIA news agency.[2] Since question topic classification is not the focus of our study, we opted for the following simple approach. We manually expanded each topic with synonymous phrases, e.g., "gay marriage" → "same-sex marriage" (on average, five phrases for each topic). A question was then considered on a topic if its lemmas contained all the lemmas of one of the topic phrases. Filtering the questions using the expanded phrases for the 19 topics resulted in 4.5 million questions.

We then sampled 54,850 questions and annotated them in two subsequent labeling tasks on the crowdsourcing platform Toloka:[3] (1) *topic aboutness* to label questions with whether they are about the controversial topic (*on topic*), contain the topic's lemmas but do not ask about the topic (*not on topic*), or are not grammatically correct questions (*ill-formed*). An example of a not-on-topic question

[2] https://ria.ru/20121221/915705250.html
[3] `https://toloka.ai/`

is “What is evolution of marketing?” which does not ask about Darwin’s theory of evolution (our controversial topic). And (2) *question type labeling* of on-topic questions into the following question types:

- *Factual questions* asking about information that most people agree on (facts), e.g., “Which countries legalized marijuana?”
- *Argumentative questions* seeking arguments or opinions for or against a topic or a statement in a question—an answer would ideally contain reasoned evidence which people might accept, reject, or doubt, e.g., “Should marijuana be legalized?”
- *Method questions* seeking a list of instructions or a description of a method to reach a goal, e.g., “How to hold a referendum on legalizing marijuana?”

These three question types differ in how widely acceptable their answers are. An answer to factual questions is a single fact that can be verified. On the other hand, a multitude of opposing and acceptable arguments exist to argumentative questions. Similarly, different lists of instructions exist for achieving a goal, which in turn differ in the required effort to follow them and their outcomes.

We conducted both annotation tasks in two steps: a pilot study to test the annotation tasks and collect quality checks and the main study. The annotation instructions for both tasks included the description of the labels as provided above and an example for each label. For topic aboutness, we provided an excerpt of the corresponding Wikipedia article that describes the topic. We split questions belonging to a topic into batches of 10 items, one of which was a quality check. We assigned three workers to each task and allocated a new worker in case one of the workers got suspended due to low annotation quality. To guarantee the quality of annotations, the tasks were conducted with a qualification test and quality checks. The qualification tests for both tasks comprised 25 pre-annotated questions on the topic “death penalty”. Workers were admitted to the annotation tasks if their accuracy exceeded 70% in the qualification test on topic aboutness and 50% on question type labeling. We suspended workers whose accuracy on the quality checks was lower than the specified threshold (70% for topic aboutness and 50% for question types).

Pilot Study We randomly sampled 120 questions from the dataset on the 19 topics (cf. Table 3.1), resulting in 2,280 questions. From these questions, 25% were used as quality checks and qualification tests and were annotated by two experts who are native Russian speakers. The rest 75% of the questions were labeled by crowd workers. The workers’ inter-annotator agreement for the topic aboutness annotations was a Krippendorff’s α = 0.55 and for question type labeling α = 0.45.

Main Study The main annotation phase covered the 52,570 questions that remained after excluding the questions used in the pilot study. Questions with perfect agreement in the pilot study were added to the quality checks. During the annotation, we iteratively expanded the set of quality checks from crowdsourced annotations with perfect agreement. We used each of the quality checks only once to ensure that workers do not memorize them.

The workers achieved an α of 0.55 on the topic aboutness task and an α of 0.49 on question type labeling. The questions on which the crowd workers achieved majority agreement amounted to 50,316 questions (92% of all questions) and were used to construct the final dataset.

Annotation Results Table 3.2 shows the distribution of the annotated questions over the topic aboutness labels and question types. The statistics show the merit of conducting the topic filtering steps as 24% of the sampled questions are not on the 19 controversial topics. The majority (64%) of the questions on controversial topics in our study look for facts, while 28% of the questions look for arguments. This indicates that people use search engines more often to look for some background information about controversial topics like factual evidence, but the share of argumentative questions is substantial.

3.2 Quantitative Question Analysis

Having crowdsourced a question dataset on controversial topics, we analyze what distinguishes argumentative questions from factual and method ones. Our analysis mainly targets four characteristics of questions that we assume to set apart argumentative questions from the other question types: question words, predictions, comparisons, and personal pronouns. To capture the characteristics in a question, we develop patterns that use surface features of questions (e.g., lemmas, part-of-speech tags, and tense information), which we extract using the mystem tagger.[4]

Question Words Question words are a strong indicator of the answer type for a question. Early research on question answering considered factual questions to start with wh-words [95] and mapped each wh-word to an entity type (e.g., "time" for `when`). Compared to wh-questions, which seek short answers, yes/no questions are statements which are converted into questions. In the context of controversial topics, we expect yes/no questions to be claims that the users have and would like to collect evidence for. On the other hand, we anticipate that questions starting with wh-words look for background knowledge about a controversial topic.

[4] `https://yandex.ru/dev/mystem/`

TABLE 3.3: The absolute and relative count of factual, argumentative, and method questions in the dataset for each characteristic in the analysis.

Characteristic	**Factual**		**Argumentative**		**Method**		**All**	
	Rel	**Abs**	**Rel**	**Abs**	**Rel**	**Abs**	**Rel**	**Abs**
Yes/No	7.2%	1,743	**13.8%**	**1,501**	0.4%	13	8.3%	3,262
Predictions	3.8%	921	**8.2%**	**892**	0.6%	19	4.8%	1,875
Comparisons	3.2%	777	**5.7%**	**625**	4.4%	130	4.0%	1,559
Personal Pronouns	0.3%	83	**3.8%**	**412**	0.43%	13	1.3%	508
Wh-words	67.0%	16,980	62.0 %	6,807	**94.3%**	**2,852**	67.7%	26,639
`why`	1.3%	325	**20.7%**	**2,253**	0.0%	1	6.6%	2,605
`how`	7.0%	1,704	7.6 %	833	**87.5%**	**2,590**	13.4%	5,274
`how much/many`	**10.5%**	**2,553**	2.1%	228	0.2%	5	7.4%	2,914
* `money`	**3.4%**	**819**	0.4%	43	0.1%	2	2.3%	907
* `people`	**2.4%**	**585**	0.7%	81	0.0%	0	1.8%	691
* `time`	**1.3%**	**322**	0.1%	14	0.0%	0	0.9%	354

Personal Pronouns We expect search engine users to refer to themselves or to an imaginary audience while formulating an argumentative question. To capture such questions, we extracted all questions whose subject is a first-person or a second-person pronoun.

Predictions One way of approaching a controversial topic is deliberation, where people try to argue for a possible course of action by predicting its consequences. We expect a subset of argumentative questions to ask for predictions that pertain to the controversial topic (e.g., “Will legalizing marijuana reduce crime?”). To extract prediction questions, we developed a pattern that looks up whether the first verb is `will` or whether it is in the future tense.

Comparisons Controversial (or argumentative) topics can also be formulated as a comparison between at least two options (e.g., death penalty vs. life imprisonment). A recent study on comparative questions asked on the web shows that more than 50% of such questions are argumentative, not factual [27]. To identify comparative questions in our dataset we apply eight regular expressions that were proposed in [30] and that were shown to classify comparative questions with a precision of 1.0.

The distribution of factual, argumentative, and method questions in the extracted questions for each characteristic is shown in Table 3.3. We also list examples of the extracted questions for each characteristic, together with the question type most associated with it in Table 3.4. By comparing the relative counts for

TABLE 3.4: Example questions for each characteristic in the analysis and their question types: factual, method, and argumentative (Arg). The questions are translated from Russian to English.

Characteristic	**Question**	**Type**
Yes/No	Is marijuana legalization possible?	Arg
Predictions	Will marijuana be legalized in Russia?	Arg
Comparisons	Should we have partial or full marijuana legalization?	Arg
Personal Pronouns	Do you think the president will legalize marijuana?	Arg
`why`	Why are people in favor of legalizing marijuana?	Arg
`how`	How to fill an amendment for marijuana?	Method
`how much money`	How much does marijuana cost?	Factual
`how many people`	How many people consume marijuana?	Factual
`how much time`	How many hours can one detect marijuana in the body?	Factual

factual and argumentative yes/no questions, we observe that they are almost twice more likely to be argumentative than factual. Wh-questions, on the other hand, cover almost the same proportion (two-thirds) of factual and argumentative questions and the majority of method questions. By analyzing the distributions for the single question words, we notice clear associations of some of them with the question types that we report in the table. As illustrated in the table, 20% of argumentative questions start with `why`, which shows that users ask explicitly for reasons when they look for arguments. A stronger association can be seen for method questions which are dominated by `how` with 87.5%. Interestingly, about 10% of factual questions look for quantities using `how much/many`. We customized the regular expression to capture different types of quantities people ask for (`money`, `people`, and `time`) by specifying synonymous verbs or nouns to the quantity type after `how much/many`. It turns out that a third of the questions that look for quantities ask about money, while about 20% ask for the count of people.

A closer look at the question type distribution for predictions shows that 8.2% of argumentative questions are written in the future tense in comparison to 4.4% for the other question types. These numbers confirm our assumption that asking for predictions is a strong indicator of argumentative questions. Personal pronouns match almost only argumentative questions, which renders personal pronouns a strong indicator of argumentative questions. Still, the very low percentage of matched argumentative questions (3.8%) shows that users formulate argumentative questions more objectively. Comparative patterns are insufficient to distinguish argumentative questions since the relative count of method questions is quite close to that of argumentative (5.7% versus 4.4%).

TABLE 3.5: F1-score for classifying questions on controversial topics into factual, method, and argumentative; in-topic and cross-topic settings.

	In-topic				Cross-topic			
Classifier	**Fact.**	**Method**	**Arg.**	**Macro**	**Fact.**	**Method**	**Arg.**	**Macro**
Random	0.44	0.13	0.31	0.29	0.43	0.12	0.30	0.28
Majority	0.78	0.00	0.00	0.26	0.78	0.00	0.00	0.26
Rule-based	0.71	0.62	0.48	0.60	0.69	0.56	0.46	0.57
Logistic regression	0.86	0.70	0.67	0.74	0.80	0.52	0.61	0.65
RuBERT	0.90	0.83	0.78	0.84	0.85	0.74	0.74	0.78

3.3 Experiments

In this section, we assess the effectiveness of automatic classifiers that map the questions in our dataset to their question types (factual, method, or argumentative). Since new controversial topics emerge all the time, a key challenge lies in generalizing beyond the 19 topics contained in the dataset.

To assess this, we conduct in-topic and cross-topic experiments on the dataset where we control for the topic differently. We use only questions that are labeled on topic in our dataset in both experiments. The in-topic experiments are conducted in a 5-fold cross-validation fashion. While sampling the folds, the dataset questions are stratified by their topics, making each fold equally cover all the topics. The cross-topic experiments, on the other hand, are conducted in a leave-one-out cross-validation fashion. Here, we use all the questions on one topic as a test set while taking the remaining questions as a training set. As evaluation metrics, we use F1-score for each of the three question types and their macro average.

Our classifier is based on RuBERT [85], which is a BERT [46] model pre-trained on the Russian Wikipedia and news articles. We feed the question to RuBERT as `[CLS] question [SEP]` and fine-tune it for two epochs with a learning rate of 2×10^{-5}.

In addition, we use four baselines: random baseline, majority baseline, a rule-based classifier, and logistic regression. The rule-based classifier relies on the insights gained from the analysis in Section 3.2, which shows a strong association between the wh-words and the three question types. The rule-based classifier categorizes a question into factual if it starts with one of the wh-words, except for `how` and `why`, for which the classifier predicts the question types method and argumentative, respectively. In case the question starts with any other word, it is classified as argumentative. The logistic regression classifier takes the count of 1-3-grams and the count of part-of-speech 1-3-grams in the question as features.

Table 3.5 shows the classification results in the in-topic and cross-topic experiments. The rule-based classifier reaches a comparable macro F1-score of 0.57 in both experiments, showing that question words are a strong indicator of the ques-

TABLE 3.6: Examples of questions in the test datasets of the cross-topic experiments which RuBERT classified wrongly.

Question	**Label**	**Prediction**
Should gays be allowed to marry?	Argumentative	Factual
How was death penalty done in the USSR?	Method	Factual

tion type regardless of the topic. RuBERT is more robust across topics than logistic regression and suffers only a drop of 0.06 macro F1-score between the two experiments in comparison to 0.09 for logistic regression. Whilst RuBERT and logistic regression perform very well on factual questions, RuBERT performs substantially better on non-factual questions.

3.3.1 Error Analysis

The results of the experiments show promising results in classifying questions on controversial topics into factual, method, and argumentative. Still, the effectiveness of RuBERT in the cross-topic setting (F1-score of 0.78) indicates a large potential to improve the classifier. To this end, we conduct an error analysis that aims at detecting systematic errors that provide insights into how to improve the approach. In the error analysis, we manually check questions in the test sets of the cross-topic experiments for which RuBERT predicts the wrong question type.

Overall, we find that the most confused question types are factual and argumentative, with 2,995 factual questions classified as argumentative and 2,683 argumentative questions classified as factual. We notice that the cause of some errors is keywords or the question tense which are correlated with factual or argumentative questions. Table 3.6 shows examples of these errors. Some keywords are often used in factual questions in the dataset (e.g., “allowed” or “approve”). RuBERT seems to rely extensively on such keywords, causing argumentative questions that use them to be classified as factual (e.g., Question 1 in Table 3.6). A similar case can be observed for questions in the past tense, which is more used in factual questions. Because of this, RuBERT tends to classify method questions in the past tense as factual (e.g., Question 2 in Table 3.6). The analysis shows that RuBERT tends to rely on surface features to predict the question type. This can be explained by the scarce context provided in the question and hints at the need to expand the question with more information about the topic.

3.4 Summary

In this section, we annotated in a crowdsourcing task a question dataset that is sampled from the Yandex query log and covers 19 controversial topics. Each of the questions is labeled with whether it is on one of 19 controversial topics, and if so, with whether it looks for a fact, a method, or arguments. The crowdsourcing study shows that the percentage of argumentative questions is high (28%), which clearly speaks for the importance of properly answering them. A comparative analysis of argumentative questions against the other question types provides first insights into their structure and properties: argumentative questions tend to ask for reasons and predictions. Experiments on the dataset show high effectiveness (F1-score of 0.78) in automatically classifying questions into argumentative, factual, or method, even on unseen topics.

Chapter 4

Generalizability of Argument Mining Approaches

Answering web queries that seek arguments requires a suitable source of arguments. The web offers the largest source of information that we know of, which guarantees a broad coverage of argumentative content. Existing approaches for mining arguments include several steps: segmenting a document into argument units, classifying their argumentative roles, and classifying their stance on a given topic.

Developing argument mining approaches to extract arguments from the web is obstructed by its different genres. This chapter starts by developing an approach to detect argument units in a genre-robust way. Since argument units might span multiple sentences or just a couple of words, we model argument units on the token level by labeling each token with regard to an adjacent or enclosing argument unit with (B)eginning, (I)nside, and (O)utside. We relabel three existing argument corpora that represent different genres with the proposed token-level annotation scheme. In in-genre and cross-genre experiments, we develop and evaluate three different machine learning models that predict the label of a token while encoding three different broadness-levels for the context.

Argument mining approaches should be effective at providing arguments to topics that are not covered by the corpora on which they are trained. Hence, the generalizability of argument mining approaches to new topics is an essential condition for retrieving arguments to web search queries. Assessing the generalizability of argument mining approaches across topics requires controlling for the topic while designing experiments. A review of existing argument corpora in Subsection 4.2.1 shows that most existing argument corpora are created without clear topic selection guidelines, and that a third of them are not labeled with topics at all. In Section 4.2, we introduce three argument topic ontologies, which are graphs, whose nodes are controversial topics that are selected by domain experts. Using these topic ontologies, we assess the proportion and distribution of controversial topics in 31 argument corpora.

Apart from corpus construction guidelines, the generalizability of an argument mining approach to new topics depends on how the task to be tackled is defined.

The last section introduces a topic-agnostic variant of stance classification, which aims to allow approaches to be less dependent on the topic. To this end, we introduce the same side stance classification task (SSSC), which is a reformulation of the stance classification task that takes as input a pair of arguments and returns whether the arguments are on the same or opposite side. We expect the SSSC task to be more topic-agnostic than stance classification since an approach for stance classification might learn topic-specific features, which makes it harder to generalize over topic. On the other hand, an approach for SSSC has to assess only the similarity between two arguments within a stance, which makes it more robust across topics.

4.1 Argument Unit Segmentation over Genre

Unit segmentation is often seen as the first task of an argument mining pipeline. It consists of splitting a text into its argumentative segments (called argument units from here on) and their non-argumentative counterparts. Afterward, the roles that the argument units play in the argumentative structure of the text as well as the relations between the units are classified. Conceptually, an argument unit may span a clause, a complete sentence, multiple sentences, or something in between. The size of the units depends on the genre of an argumentative text but can also vary within a text. This makes unit segmentation a very challenging task.

As detailed in Section 2.3.1, much existing research on argument mining has skipped the segmentation step, assuming argument units to be given. For applications such as argument retrieval, however, automatic segmentation is obligatory. Different approaches have been presented that deal with unit segmentation of argumentative essays: Persing and Ng [124] rely on handcrafted rules based on the parse tree of a sentence to identify segments; Stab [154] use sequence modeling based on sophisticated features to classify the argumentativeness of each single word based on its surrounding words; and Eger et al. [50] employ a deep learning architecture that uses different features to do the same classification based on the entire essay. So far, however, it is neither clear what the best segmentation approach is, nor how different features and models generalize across genres of argumentative texts.

In this section, we aim at developing an effective unit segmentation approach and assessing its generalizability across genres. We follow the outlined work in tackling unit segmentation as a token-level classification task (Section 4.1.2). To capture the context around each token, we analyze different semantic, syntactic, structural, and pragmatic feature types, and we compare three fundamental machine learning techniques based on these features: standard feature-based classification realized as a support vector machine (SVM; [41]), sequence modeling realized as linear-chain conditional random field (CRF; [86]), and a deep learn-

TABLE 4.1: Number of documents, tokens per label, and average tokens per document per corpus and part. Tokens in the three corpora are labeled with Arg-B, Arg-I, and Arg-O, which stand for the beginning, inside, and outside of an argument unit.

			Number of tokens				
Corpus	**Part**	**# Documents**	Arg-B	Arg-I	Arg-O	Total	Average
Essays	Train	322	4,823	75,621	35,323	115,767	359.5
	Test	80	1,266	18,790	8,699	28,755	359.4
	Total	402	6,089	94,411	44,022	144,522	359.5
Editorials	Train	240	11,323	202,279	17,227	230,829	961.8
	Test	60	2,811	49,102	4,622	56,535	942.3
	Total	300	14,234	251,381	21,849	287,364	957.9
Web Discourse	Train	272	905	32,093	36,731	69,729	256.4
	Test	68	224	7,949	8,083	16,256	239.1
	Total	340	1,129	40,042	44,814	85,985	252.9

ing approach realized as a bidirectional long short-term memory (Bi-LSTM; [63]). These models correspond to increasingly complex levels of modeling context: The SVM considers only the current token, resulting in an isolated classification for each word. The CRF is additionally able to consider the preceding classifications. The Bi-LSTM, finally, can exploit all words and classifications before and after the current word.

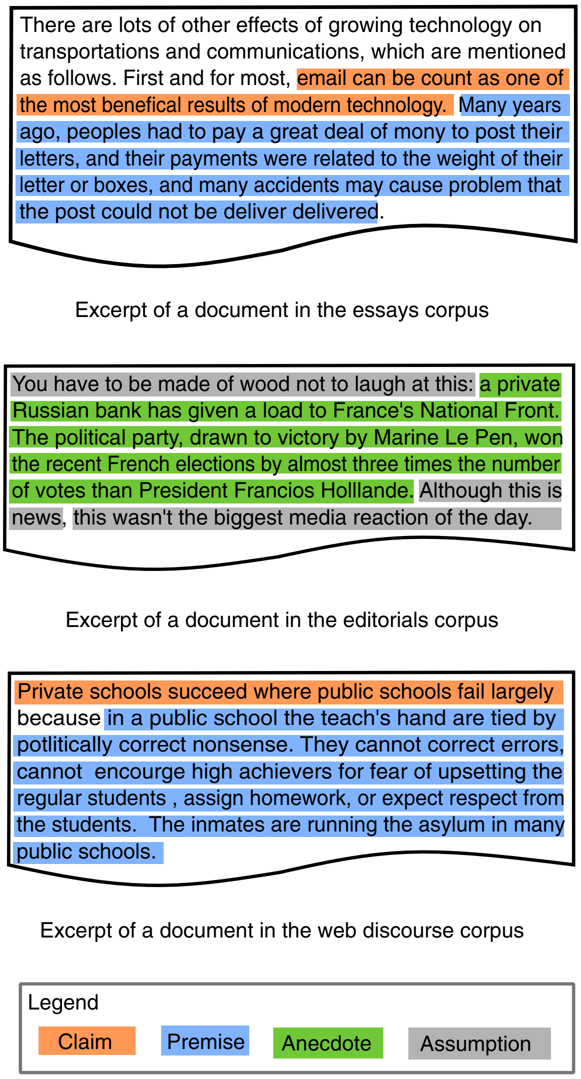
We evaluate the models on and across three existing argumentation corpora, each representing a different genre (Section 4.1.1): the Essays corpus of Stab [154], the Editorials corpus of Al-Khatib et al. [10], and the Web Discourse corpus of Habernal and Gurevych [68]. All combinations of training and test genre are considered for these corpora, resulting in nine experiments.

4.1.1 Data

This study uses three different corpora from different genres to evaluate the models that we developed to segment argument units. We detail each corpus below, give an overview in Table 4.1, and provide example excerpts in Figure 4.1).

Essays The Argument Annotated Essays Corpus [148, 150] includes 402 argumentative essays from essayforum.com, written by students. All essays are segmented by three expert annotators into argument units (major claims, claims, and premises) and non-argumentative parts. Each such argument unit covers an entire sentence or less. The essays are on average 359.5 tokens long, with 70% of tokens being part of an argument unit.[1] We employ the test-training split provided by the

[1] Percentage of tokens that are part of an argument unit is calculated from Table 4.1 as (Arg-B + Arg-I)/Total



There are lots of other effects of growing technology on transportations and communications, which are mentioned as follows. First and for most, email can be count as one of the most benefical results of modern technology. Many years ago, peoples had to pay a great deal of mony to post their letters, and their payments were related to the weight of their letter or boxes, and many accidents may cause problem that the post could not be deliver delivered.

Excerpt of a document in the essays corpus

You have to be made of wood not to laugh at this: a private Russian bank has given a load to France's National Front. The political party, drawn to victory by Marine Le Pen, won the recent French elections by almost three times the number of votes than President Francios Holllande. Although this is news, this wasn't the biggest media reaction of the day.

Excerpt of a document in the editorials corpus

Private schools succeed where public schools fail largely because in a public school the teach's hand are tied by potlitically correct nonsense. They cannot correct errors, cannot encourge high achievers for fear of upsetting the regular students , assign homework, or expect respect from the students. The inmates are running the asylum in many public schools.

Excerpt of a document in the web discourse corpus

FIGURE 4.1: Excerpts of three documents for the Essays, Editorials, and Web Discourse corpora. Each excerpt is highlighted with argument units as annotated in the original corpus.

authors.

Editorials The Webis-Editorials-16 corpus [10] consists of 300 news editorials from the three online news portals Al Jazeera, Fox News, and The Guardian. Prior to the annotation process, the corpus was pre-segmented based on clauses. After that, three annotators performed the final segmentation by merging segments and dividing argument units (common ground, assumption, anecdote, testimony, statistics, and others) from non-argumentative parts. The annotation guidelines define a unit as a segment that spans a proposition (or two or more interwoven propositions) stated by the author to discuss, directly or indirectly, his or her thesis. This corpus contains the longest documents among the three studied corpora, with an average of 957.9 tokens. The editorials are mainly argumentative, with 92% of the tokens in the corpus being part of an argument unit. We employ the test-training split provided by the authors.

Web Discourse The Argument Annotated User-Generated Web Discourse corpus [71] contains 340 user comments, forum posts, blogs, and newspaper articles, and they are annotated according to a modified version of Toulmin's model [162]. In this corpus, argument units (premise, claim, rebuttal, refutation, and backing) can be arbitrary text spans. Because of this, argument units are on average much longer than in the other two corpora: 36.5 tokens compared to 16.5 tokens (Essays) and 18.7 tokens (Editorials).[2] The texts are relatively short (252.9 tokens on average) and contain many non-argumentative parts: only 48% of the tokens are part of an argument unit. Since the authors do not provide a test-training split, we randomly split the corpus into a training set (80%) and a test set (20%), similar to the other corpora.

The three corpora vary in terms of how arguments are actually annotated in the contained documents. Following the approach of Stab [154], we converted all documents into BIO format, where each token in the documents is labeled according to the segment it belongs to as *Arg-B* (the first token of an argument unit), *Arg-I* (token inside an argument unit), or *Arg-O* (not in argument unit).

4.1.2 Approach

In line with recent literature, we address unit segmentation as a token labeling problem. Specifically, we classify each token in an input document into *Arg-B*, *Arg-I*, or *Arg-O*. We explore the effectiveness of semantic, syntactic, structural, and pragmatic features when capturing tokens separately, along with their neighbors, or along with the entire text. In the following, we detail each set of features and each of the three machine learning models we use, where each model reflects a different broadness of context that is used to classify the tokens. To demonstrate the strengths and weaknesses of the models, we encode the features as analog as possible in each model. However, some variations are necessary due to differences in the way the models utilize the features.

Features

For every token, we extract the following semantic, syntactic, structural, and pragmatic features.

[2] Average length of argument units is calculated from Table 4.1 as $(\text{Arg-B} + \text{Arg-I})/\text{Arg-B}$

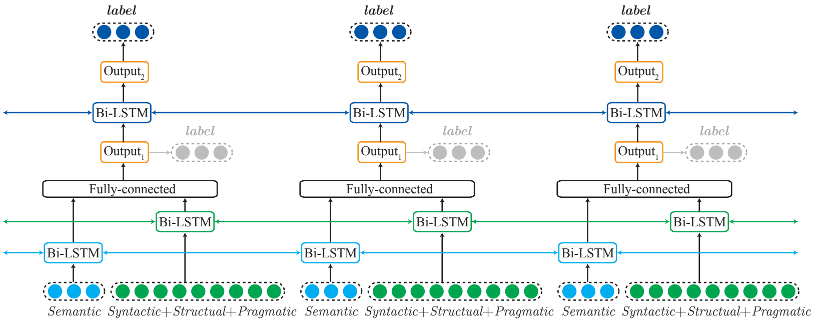


FIGURE 4.2: The proposed neural network structure with the input feature vectors for three tokens at the bottom. The labels at $Ouput_1$ are estimated without considering label dependency and are not used; instead, we report the results for $Output_2$, which considers this dependency.

Semantic Features Semantic features capture the meaning of tokens. This work employs the simple but often effective way of representing meaning by using the occurrence of each token as a feature (bag-of-words). We also tested word embeddings [123] as semantic features but found that they performed worse for all but the neural network models (cf. Table 4.2).

Syntactic Features The syntactic features we employ capture the role of a token in a sentence or argument unit. We employ standard part-of-speech tags (POS) as produced by the Stanford POS tagger [163] for this feature set with one binary feature for each POS-tag.

Structural Features Structural features capture the congruence of argument units with sentences, clauses, or phrases. We employ the Stanford parser [82] to identify sentences, clauses, and phrases in the text and represent them with token labels. In particular, we use three binary features for each token and structural level (sentence, clause, phrase) that are "1" when the token is at the beginning, within, or at the end of such a structural span, respectively.

Pragmatic Features Pragmatic features capture the effects the author of a text intended to have on the reader. We use lists of discourse markers compiled from the Penn Discourse Treebank [130] and [154] to identify such markers in the text. The lists by Stab are specifically created for detecting argument units. For each token and discourse marker, we use five binary features that are "1" when the token is before the marker, the beginning of the marker, inside a multi-token marker, the last token of a multi-token marker, or after the marker in the sentence, respectively.

Models

We resort to three common machine learning models in order to capture an increasing amount of context for the token labeling: a support vector machine (SVM), a conditional random field (CRF), and a bidirectional long short-term memory (Bi-LSTM). To provide a comparison to results from related work, we reimplement the method of Stab [154] and use it as a baseline.

Reimplementation The approach of Stab [154] is a CRF model that is specifically developed for the Essays corpus. Since the license of the original implementation prohibits the author from giving us access to the code, we fully reimplemented his approach. Analogously to Stab, we employ the CRFSuite [111] with the averaged perceptron method [39]. For the reimplementation, we use the exact feature sets described by Stab: structural, syntactic, lexSyn, and prob. Our reimplementation achieves a slightly worse F1-score of 82.7 compared to the reported 86.7 for unit segmentation (Table 4.2). We attribute this difference to implementation details in the employed features.

SVM We employ an SVM model in terms of a standard feature-based classifier that labels each consecutive token independently, disregarding the token's context. In other words, features of neighboring tokens are not considered by the SVM. Accordingly, this model does not capture the transition between labels, as well.

CRF We implement a CRF model to capture the context around the token for labeling the token. For labeling, the linear-chain CRF that we use considers the labels and features of the surrounding tokens within a certain window, which we chose to be of size five for our experiments. We use the same framework and method for the reimplementation.

Since CRFs explicitly capture the local context of a token, we simplify the pragmatic features for this model and use only binary features to indicate whether the token is at the beginning, inside, at the end, or outside of a discourse marker.

Bi-LSTM We also build a neural network to capture the entire text as context. Instead of using the tokens directly as semantic features, we use their word embedding [123] as it is common for neural networks. In particular, we use the standard pre-trained embedding by Pennington et al. [123], which has a dimensionality of 300.

We now explain the architecture of our model, as illustrated in Figure 4.2. From bottom to top, we first feed the features into bidirectional LSTMs [63]. We feed the semantic features into a separate Bi-LSTM to be able to use a different kernel for this dense feature vector than for the sparse feature vectors. The output of these two Bi-LSTMs is then concatenated and fed into a fully-connected layer.

To model label dependencies, we add another Bi-LSTM and another output layer. Both output layers are softmax layers, and they are trained to fit the labels of the tokens. We only use the result of the second output layer, though. As shown in Section 4.1.3, the second output layer does indeed better capture the sequential relationship of labels.

4.1.3 Experiments

Using the three corpora as detailed in Section 4.1.1, we conduct in-genre and cross-genre experiments to answer our research question. For both in-genre and cross-genre experiments, we use the training set for training the model and the test set for its evaluation. For each experiment, we test all four different feature sets both in isolation and in combination. We report the macro F1-score as an evaluation measure for comparison to related work, and since we consider all three classes (*Arg-B*, *Arg-I*, and *Arg-O*) to be equally important.

Table 4.2 lists the macro F1-scores for all combinations of features and models, test set (first row), and training set (second row). The Table also shows the results of our reimplementation of Stab's approach for all combinations of test and training sets.

Comparison to Stab [154] To put our results into context, we compare our methods to the approach of Stab [154]. For this purpose, we randomly split the test set of the Essays corpus into five equally-sized subsets and use the student's t-test to compare the F1-scores of our best-performing method on each subset with the result of Stab. We find that our Bi-LSTM approach achieves a significantly better F1-score (88.54 versus 86.70 with p-value < 0.001).

Furthermore, although the results of our reimplementation of Stab's approach are lower than his reported results, our own CRF approach performs comparably well in almost all cases using only simple linguistic features.

Improvement by Second Output Layer A side-effect of classifying BIO labels for each token is that two consecutive tokens can be labeled as *Arg-O* and *Arg-I*, which is not reasonable as this would correspond to a unit without a beginning. Without the second output layer $Output_2$, our neural network method produced about 400 of such unreasonable pairs. However, when we added the second output layer, this number dropped by half to 200 pairs. While the effect on the F1-score is small, using the second output layer, therefore, produces more comprehensible results. We thus only report the results with $Output_2$.

TABLE 4.2: The in-genre (gray background) and cross-genre F1-scores of the three models (SVM, CRF, and Bi-LSTM), using each of the four feature types (semantic, syntactic, structural, and pragmatic) in isolation as well as in combination (all). For each column, the highest F1-score is marked in bold. The bottom line shows the effectiveness of our reimplementation of the approach of Stab [154].

		Test on Essays			**Test on Editorials**			**Test on Web Discourse**		
Features	**Models**	Essays	Editorials	Web Dis.	Essays	Editorials	Web Dis.	Essays	Editorials	Web Dis.
Semantic	SVM	53.42	40.89	28.89	50.00	53.96	16.20	31.71	26.58	33.34
	CRF	76.56	53.06	26.31	66.30	78.90	8.48	37.51	37.25	42.53
	Bi-LSTM	87.91	**57.11**	36.00	60.70	81.56	24.63	**41.29**	36.44	**54.98**
Syntactic	SVM	49.66	36.14	26.45	49.98	51.36	14.32	28.44	25.33	31.93
	CRF	66.79	48.40	15.48	68.30	76.74	5.05	34.73	38.13	24.25
	Bi-LSTM	83.10	55.70	21.65	64.92	80.35	15.28	36.58	37.40	43.02
Structural	SVM	41.19	36.14	26.45	49.53	77.71	5.96	27.97	37.98	27.52
	CRF	60.12	48.41	15.48	68.96	77.55	5.68	34.64	**38.30**	22.51
	Bi-LSTM	69.77	48.63	**41.19**	61.54	79.62	**38.08**	35.46	37.75	39.51
Pragmatic	SVM	38.75	28.65	30.09	31.33	33.02	22.38	30.85	22.24	35.59
	CRF	40.15	31.66	15.48	37.06	40.20	5.02	24.30	30.30	23.70
	Bi-LSTM	76.47	54.72	15.24	57.66	75.31	5.24	34.88	36.68	22.76
All	SVM	61.40	50.88	31.26	58.84	79.89	22.55	39.14	37.42	42.76
	CRF	79.15	52.50	21.74	**69.80**	81.97	8.00	37.09	37.63	37.74
	Bi-LSTM	**88.54**	**57.11**	36.97	60.69	**84.11**	20.85	39.78	36.56	54.51
Reimplementation		82.70	52.00	20.00	67.00	78.00	6.00	31.66	37.30	49.00

TABLE 4.3: Pearson correlation between argument unit boundaries and structural features. Values range from -1.00 (total negative correlation) to 1.00 (total positive correlation). Absolute values above or equal to 0.40 can be seen as moderately correlated and are marked in bold.

		Sentence			**Clause**			**Phrase**		
Corpus	**Label**	B	I	E	B	I	E	B	I	E
Essays	Arg-B	0.30	-0.19	-0.05	0.23	-0.13	-0.06	0.04	0.04	-0.08
	Arg-I	-0.30	**0.44**	-0.30	-0.23	0.34	-0.22	0.04	0.03	-0.08
	Arg-O	0.18	-0.37	0.33	0.14	-0.29	0.25	-0.06	-0.04	0.11
Editorials	Arg-B	**0.75**	**-0.51**	-0.05	**0.57**	-0.38	-0.07	0.15	-0.09	-0.09
	Arg-I	**-0.53**	**0.74**	**0.48**	**-0.44**	**0.58**	-0.33	0.02	0.12	0.11
	Arg-O	0.05	**-0.50**	**0.64**	0.09	**-0.41**	**0.47**	-0.10	-0.09	0.21
Web Discourse	Arg-B	**0.48**	-0.33	-0.03	0.32	-0.22	-0.04	0.10	-0.06	-0.05
	Arg-I	-0.12	0.09	0.00	-0.09	0.07	0.00	-0.02	0.01	0.01
	Arg-O	0.18	0.01	-0.01	0.01	0.01	0.08	0.00	0.00	0.00

4.1.4 Discussion

Given our experimental results, we come back to the research question we initially raised and then turn our heads to ongoing research. Our study aims to provide insights into what approach works best for argument unit segmentation, and how well such an approach generalizes over genres. In the following, we compare the features and models in the light of the experimental results and discuss their robustness across genres.

Features According to the results of the in-genre experiments, the semantic features are the most effective ones. The models employing these features achieve the highest F1-scores, except for the SVM on editorials, where structural features perform better. However, there is no feature type that dominates the cross-genre experiments. At least, the structural features seem rather robust when the training and test sets are from different genres.

While the results of the semantic features across argumentative essays and editorials—two genres that are comparably similar—remain high, the performance of the models employing them dramatically drops when tested on Web Discourse after training on either of the other. The intuitive explanation for this decrease in the genre transfer is that important content words are genre-specific. Thus, the learned knowledge from one genre cannot be transferred to other genres directly. In contrast, the structural features capture more general properties of argumentative text, which is why we can use them more reliably in other genres.

As shown in Table 4.3, the sentence, clause, and phrase boundaries correlate with the boundaries of argument units. Especially in the Editorials corpus, the boundaries of sentences and clauses show high Pearson coefficients. This reveals why we can still achieve reasonable performance when the training and test sets differ considerably.

Models Comparing the different models, the SVM performs worst in most experiments. This is not surprising because the SVM model we used utilizes local information only. In a few cases, however, the SVM performed better than the other models, e.g., when evaluating pragmatic features on argumentative essays that were learned on web discourse. One reason may be that such features rather have local relevance. As a matter of fact, adding knowledge from previous and preceding tokens will add noise to a model rather than be beneficial.

Overall, the models employing sequential features turn out to be stronger. Among them, the Bi-LSTM model mostly achieves the best results regardless of the genre or the features. This suggests that context information from the tokens around a token to be classified is generally useful. In addition, using neural networks seems to be a better choice to encode these features. Additionally, another advantage of using a Bi-LSTM is that the Bi-LSTM can utilize all features of tokens from the beginning to the end of the document. This allows the model to capture long-distance dependencies. For a CRF, such dependencies are hard to encode, requiring to increase the complexity of the model dramatically and thus making the problem intractable.

Genres From the results and the previous discussion, we conclude that our structural features (capturing the boundaries of phrases, clauses, and sentences) and the Bi-LSTM model are the most genre-robust. Other features, especially semantic features, tend to be more genre-dependent. The ability to model long-distance dependencies and a more advanced feature encoding indicate why the Bi-LSTM apparently learns more general, less genre-specific features of the given argumentative texts.

Major Challenges of Unit Segmentation

The effectiveness loss in the genre transfer suggests that the notion of an argument unit is not entirely the same across argumentative text corpora. This hypothesis is supported by the high variance in the size of argument units, ranging from clause-like segments [10] to partly multiple sentences [136]. At the same time, it seems reasonable to assume that there is a common concept behind that connects the different notions of argument units, and that distinguishes them from other types of segments. Under this assumption, a general question arises that we see as fundamental in research on unit segmentation:

Open Question about Argument Units: What makes argument units different from other syntactic units, and at what point do they deviate?

While it is possible to study this question based on a matching of the argument units and syntactic units in a given dataset, a generally satisfying answer might not exist because we expect the segmentation into argument units to be task-specific to some extent. Similar observations have been made for elementary discourse units [157]. In case of argument units, some annotations, for example, model the hierarchical structure of a text primarily [154], while others aim to capture self-contained evidence [136]. Even for a given task, however, unit segmentation remains challenging, though, as underlined by the limited effectiveness we observed in some cases. As a result, it is a topic of ongoing discussion in the community. This brings up another question:

Open Question in Unit Segmentation: What knowledge is needed to effectively perform unit segmentation?

In particular, it has been discussed controversially in the community as to whether unit segmentation should actually be tackled as the first step of argument mining. When tackled first, no knowledge about the main claims of an argumentation, the applied reasoning, and similar is given, making the feasibility of distinguishing argumentative from non-argumentative parts doubtful. Of course, other orderings might lead to analog problems, which would then suggest to jointly approach the different steps.

4.2 Topic Bias in Argument Corpora

Topics play a central role in argumentation since they define the matter of contention and the needed context to construct and evaluate arguments. The context around a subject matter defines, for example, the main actors involved in the contention. For the controversial topic “legalizing marijuana”, for example, the actors can be “minors” who should be protected from using it and “drug dealers” who have economic benefits from selling it. A topic constraints or guides the persuasion strategies that can be used in an argument about it [167]. A persuasion strategy that is dependent on a topic can frame the topic by highlighting a specific aspect (e.g., “we should legalize marijuana since it has crucial health effects”, here, the frame is “health”). Another persuasion strategy is to use rhetorical figures (e.g., analogies such as “We do not ban alcohol and smoking even though they are both harmful.”)

To guarantee that computational argumentation approaches generalize over topics, researchers should control for the topic while developing them. This is especially important for supervised approaches since they can capture topic-specific features while learning how to extract arguments. Reuver et al. [134] and Jakobsen et al. [78] show that transformer-based classifiers usually capture topic-specific features (e.g., the word “kill” for the topic “abortion”).

TABLE 4.4: Survey of argument corpora indicating data source, unit granularity, and size in terms of units and topics (if the authors remarked on it). The unit granularity is the one in the corpus files, with argument pairs treated as two units and using the best context-preserving unit in case the corpus features multiple granularities. The grouping indicates our presumed topic selection directive. Selected implies a preference made by the authors regarding the topics, while source-driven implies that the choice of the topics is decided by the source. Experiments (Exp.) is the count of papers that use the corpus in an experiment among those papers the cite the paper describing the corpus.

Corpus	**Authors**	**Source**	**Unit granularity**	**Units**	**Topics**	**Exp.**
		Selected				
WebDiscourse	[71]	Web	Document	340	6	7
UKP Sentential	[151]	Web	Argument	25,492	8	13
Internet Argument Corpus v2	[1]	Web	Discussion	16,555	19	18
UKP Aspect	[133]	Web	Argument pair	3,595	28	3
Key Point Analysis	[21]	Wikipedia	Argument	24,093	28	2
Argumentative Sentences	[53]	Wikipedia	Arguments	700	20	1
Claim and Evidence 1	[2]	Wikipedia	Wikipedia article	315	33	18
Claim Stance	[19]	Wikipedia	Argument Unit	2,394	55	10
Claim and Evidence 2	[136]	Wikipedia	Wikipedia article	547	58	12
Evidence Quality	[62]	Wikipedia	Argument pair	5,697	69	1
Claim Sentence Search	[93]	Wikipedia	Argument unit	1,492,077	150	3
Evidence Sentences	[142]	Wikipedia	Argument unit	5,783	118	5
Evidence Sentences 2	[51]	Wikipedia	Argument unit	29,429	221	3
Multilingual Argument Mining	[161]	Wikipedia	Argument unit	65,708	347	2
Arguing Subjectivity	[40]	Editorials	Editorial/blog	84	1	1
COMARG	[26]	Debate portals	Argument pair	2,298	2	3
Argument Facet Similarity	[103]	Debate portals	Argument	6,188	3	8
Ideological Debates Reasons	[76]	Debate portals	Argument	4,903	4	10
Webis-debate-16	[9]	Debate portals	Debate	445	14	3
UKPConvArg1	[69]	Debate portals	Argument pair	11,650	16	10
UKPConvArg2	[70]	Debate portals	Argument pair	9,111	16	3
Political Argumentation	[99]	Debating	Argument pair	1,462	5	3
Record Debating Dataset 2	[102]	Debating	Speech	200	50	5
Record Debating Dataset 4	[112]	Debating	Speech	200	50	1
Record Debating Dataset 3	[88]	Debating	Speech	400	199	1
Record Debating Dataset 5	[113]	Debating	Speech	3,562	397	1
ICLE Essay Scoring	[126]	Essays	Essay	1,000	10	11
Micro Text v1	[122]	Essays	Essay	112	18	7
Micro Text v2	[145]	Essays	Essay	171	35	1
Sci-arg	[87]	Scientific papers	Paper	40	1	3
Claim Generation	[65]	Generated text	Argument Unit	2,839	136	1

Table 4.4 (continued).

Corpus	Authors	Source	Unit granularity	Units	Topics	Exp.
Source-driven: greedy within a time-span						
AIFdb	[22]	Web	Argument unit	67,408	n/a	7
ChangeMyView	[158]	Discussion forum	Post/comment	14,066	n/a	21
Intelligence Squared Debates	[182]	Debate portals	Debate	108	n/a	3
Args-me	[6]	Debate portals	Argument	387,692	n/a	3
Kialo	[81]	Debate portals	Argument unit	331,684	n/a	3
DebateSum	[140]	Debating	Debate	187,386	n/a	1
USElecDeb60To16	[73]	Debating	Debate	42	n/a	1
Political Speech	[96]	Debating	Argument unit	152	n/a	1
Source-driven: sampled						
GAQCorpus	[110]	Web	Argument	6,424	n/a	1
Editorials	[10]	Editorials	Editorial	300	n/a	8
IDebate Persuasiveness	[125]	Debate portals	Argument	1,205	n/a	1
Argument Annotated Essays	[150]	Essays	Essay	402	n/a	28
E-rulemaking	[118]	Discussion forum	Argument	731	n/a	3
ECHR	[129]	Law	Argument	743	n/a	1

This section starts with a review of 45 argument corpora in terms of how many topics they cover and how these topics are chosen. The review shows that most argument corpora are created without commenting on topic selection, and 14 argument corpora lack any topic labeling. Another finding is that researchers conduct more experiments on argument corpora with few topic labels. Constructing argument corpora and designing experiments in computational argumentation should start with selecting topics from accepted sources of controversial topics. Since computational argumentation tasks are topic-dependent, conclusions on the generalizability of an approach can only be drawn after carefully controlling for the topic. To function as an accepted controversial topics space, we acquire and introduce three authoritative sources of controversial topics for arguments that are organized as topic ontologies. A topic ontology is a directed graph whose nodes are topics and relations encode “is part of” relations. Using the ontologies, we analyze the coverage of 31 argument corpora that are provided with topic labels. We assess the coverage of argument corpora by computing the proportion of ontology topics that is covered by 31 argument corpora and the distribution of the corpora topic labels in the ontologies.

4.2.1 Topic Selection Approaches in Existing Argument Corpora

Table 4.4 lists our review of all corpora related to argumentation, which are published by 2020. The corpora are listed together with the number of topics each corpus covers, its source, the granularity of the corpus, its size, and its associated publication. We also analyzed how many experiments were carried out using them to date. The only work where the researchers justify their topic selection is conducted by Habernal et al. [72], who chose six topics (homeschooling, public versus private schools, redshirting, single-sex education, prayers in schools, sex education, and mainstreaming) to focus on education-related topics. Still, the researchers do not mention where these topics come from. Stab et al. [152] mention the source of topics for the corpora they created, which are two lists of controversial topics: an online library and a debate portal (ProCon.org). Except for these two cases, researchers do not justify their choice of topics nor use a selection or sampling criterion while choosing the topics.

Our review shows three different ways of how researchers choose topics while constructing argument corpora:

1. Manual Selection: topics are manually selected or defined.
2. Source-driven (greedy within time span): either an argument source is exploited in its entirety or a subset is taken based on time.
3. Source-driven (sampled): a sample of a specific source is taken without regard to topics.

A manual selection of topics might incur biases toward certain known clichés or topics which might be frequently picked by researchers (e.g., "abortion"). The availability of topic labels allows us to assess the topic distribution of corpora whose topics are manually selected. On the other hand, argument corpora created with source-driven selection approaches do not include topic labels. The topics covered by these corpora come from the arguments' source, which might follow the source's preference of topics.

The unguided topic selection while constructing argument corpora can affect the generalizability of the experiments conducted on them. We count how many experiments have been reported up to April 2020 on each of the corpora by collecting the scientific papers referring to a corpus as per Google Scholar. From all referring papers, we count those papers that use the corpus in an experiment. As shown in Table 4.4, researchers tend to pick corpora with fewer topics more often than those corpora with larger amounts of topics. All in all, researchers conducted 82 experiments on argument corpora with no clearly defined topic selection directive. The tendency of researchers to choose small corpora might affect the generalizability of their approaches and the validity of their findings.

4.2.2 Acquiring Authoritative Argument Topic Ontologies

The topic dependence of computational argumentation tasks makes topic selection a crucial step while constructing argument corpora. Most existing argument cor-

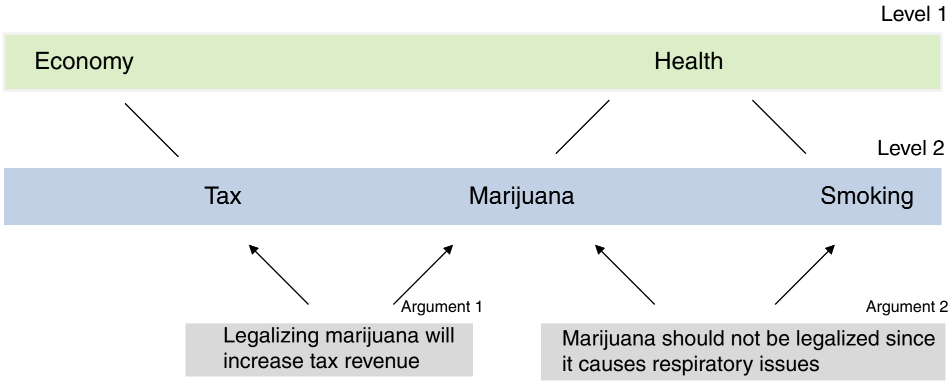

Level 1
Economy
Health
Level 2
Tax
Marijuana
Smoking
Argument 1
Legalizing marijuana will increase tax revenue
Argument 2
Marijuana should not be legalized since it causes respiratory issues

FIGURE 4.3: Example of categorizing arguments to topics of a two-level ontology. Arguments categorized as about a Level-2 topic also pertain to its Level-1 upper-topics.

pora are constructed starting from topics that are chosen or defined by researchers. For web search scenarios ensuring a representative topic coverage of users' information needs is needed.

Controversial topics do not exist in isolation and are characterized by a hierarchical structure. For example, the topic "Affordable Care Act", which is the only topic chosen by Conard et al. [40] to construct Argument Subjectivity, is a subtopic of "health care reform in the U.S.A". This more general topic includes other topics, for example, "American Health Care Act".[3] Keeping topic structure while constructing argument corpora guarantees generalizability since computational approaches are likely to generalize more across similar topics.

Topic ontologies provide a standard way to define topics and their relations. A topic ontology is modeled using a directed acyclic graph, where nodes correspond to topics and edges represent "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 other topics, continuing recursively with each lower level of subtopics. Figure 4.3 shows an example of a two-level topic ontology with two arguments categorized within it.

The identification of the topics to be included in an argument topic ontology, 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 relevant topic ontologies, which comprise a wide selection of important argumentative topics.

[3] "Affordable Care Act" and "American Health Care Act" are known colloquially as "Obamacare" and "Trumpcare" respectively.

World Economic Forum (WEF) The World Economic Forum is a not-for-profit foundation that coordinates efforts from leading organizations to confront economic and societal issues on a global scale. Strategic Intelligence is a platform initiated by the WEF that aims at informing decision-makers about important controversial topics, for example, artificial intelligence and climate change. Topics are categorized further into 4 to 9 subtopics. Domain experts curate a stream of relevant news articles for each topic and tag it with one of its subtopics.

Wikipedia A neutral point of view is the first principle of how content should be contributed to Wikipedia. Still, many topics of public interest are controversial and continuously contended. Such topics are usually a source of vandalism and edit wars [179]. Wikipedia maintains a list of such controversial articles to highlight where special care is needed.[4] The topics are grouped into 14 topics (e.g., environment and philosophy) and 4 to 176 subtopics (e.g., creationism and pollution). Omitted is the "people" topic and articles on countries; their controversiality is not universal.

Debatepedia Debatepedia's goal is to create an encyclopedia of debates that are organized as pro and con arguments. A list of 89 topics helps visitors to browse the debates. Debates on Debatepedia are contributed by anonymous web users, which makes the covered topics easily accessible. Topics in Debatepedia tend to address issues of the Western culture. For example, the topic "United States" covers 306 debates while "third world" covers 12 debates. The project is no longer actively maintained, rendering its ontology outdated. Still, the debate portal operated from 2007 to 2019, which largely overlaps with the starting time of argument mining research. While Debatepedia is no longer available online, it is archived on the Internet Archive and can be accessed through Wayback Machine.[5]

The three ontologies vary in terms of comprehensiveness and granularity. While Debatepedia covers issues more related to the U.S.A, the World Economic Forum covers issues of a wide range of countries (e.g., "Bangladesh", "the U.A.E", and "Brazil"). The World Economic Forum covers more fine-granular topics (e.g., "age-friendly infrastructure" and "Turkish monetary policy"). In comparison, Debatepedia covers more generic topics (e.g., "politics", "bans", and "religion"). In terms of covered topics, the World Economic Forum has a clear focus on the economy, covering topics such as energy, employment, infrastructure, investment, risks, and innovation. In contrast, Wikipedia and Debatepedia cover a wide range of topics, including politics, economy, law, etc.

The three ontologies are publicly accessible on the web, and two of them (Wikipedia and the WEF) are actively maintained and updated. A key task as-

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

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

TABLE 4.5: The number of topics and topic statistics for each level of the three ontologies: the World Economic Forum (WEF), Wikipedia, and Debatepedia. For each topic, we list the minimum, average, and maximum count of authors who contributed documents to the topic. In addition, we include the minimum, average, and maximum count of categorized documents and the tokens of these documents per topic. Documents with unknown authors are ignored for the author statistics.

Topic ontology	**Topics**	**Topic statistics**								
		Authors			Documents			Tokens		
		Min	Mean	Max	Min	Mean	Max	Min	Mean	Max
WEF Level 1	137	1	334.1	1,787	3	940.7	4,080	945	490,576.6	2,950,615
WEF Level 2	822	1	216.8	1,300	1	550.3	3,337	251	310,229.7	2,292,371
Wikipedia Level 1	14	9,124	78,013.7	225,623	11	68.0	172	26,977	339,088.0	1,198,813
Wikipedia Level 2	748	9	1,929.5	13,810	1	1.0	1	6	6,149.1	22,116
Debatepedia	89	12	145.0	682	7	61.7	306	2,987	84,787.6	475,033

sociated with every topic ontology is to categorize a given document into it. Having just a short string 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 wiki articles is available, as are the associated debates for Debatepedia. Documents categorization into the three ontologies is not mutually exclusive, i.e., a document can be categorized under multiple topics at the same time. For example, a Wikipedia article on “abortion” is listed under “science” and “sexuality” in Wikipedia.

We crawled the three ontologies by extracting the topics and the alongside categorized articles. Articles that are categorized into Level-2 topics are propagated up to their respective Level-1 topics. A topic description comprising one paragraph was extracted for each topic. Table 4.5 shows the large differences between the ontologies. The WEF ontology contains the most topics and links the most documents, which contain the most tokens overall. The topics at Wikipedia Level 2 are just linked to a single article each, so every topic’s amount of text is smaller. Wikipedia contains the highest count of authors, with an average of 1,929.5 authors per article. The number of authors reflects the number of editors for each topic.

4.2.3 Aligning Corpora Topic Labels to Topic Ontologies

To assess the topic coverage of the argument corpora in light of the three ontologies, we map the topic labels of those corpora, providing them with their matching ontology topics. The topic labels are first normalized by removing clichés and stance-taking language. Second, the normalized topic labels are used as queries to retrieve candidate topics from an ontology using the pre-categorized documents that come with each ontology. Finally, the candidate topics were labeled with

TABLE 4.6: Examples of topic labels in the 31 preprocessed corpora and their normalized form.

Topic label	Type	Normalized topic label	Corpus
Abortion	Concept	abortion	Claim Sentence Search
Pro Choice vs. Pro Life	Comparison	pro choice vs pro life	UKPConvArg1
Ban Abortions	Imperative	abortion	Record Debating Dataset 5
Should parents use spanking	Question	spanking	UKPConvArg1
This house would ban partial-birth abortions	Motion	partial birth abortion	Claim Evidence 2
Crime does not pay	Conclusion	crime does not pay	ICLE Essay Scoring

whether they are the upper-topic of the queried topic label.

Topic Label Normalization

Table 4.4 lists 31 argument corpora that provide a total of 2,117 topic labels. They are concise descriptions that have been provided by the corpus authors. The labels follow the style of the genre of the respective corpus: In argumentative essays, for instance, topics are usually thesis statements, while Wikipedia-derived corpora use article titles, and the topics of debate corpora include motions such as “this house should”. Often, topic labels express a stance towards a target issue, e.g., “ban abortion”. Six types of topic labels can be distinguished: concept, comparison of concepts, motion, conclusion, 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 748 unique topic labels.

Mapping Topic Labels to Ontology Topics

We map the normalized topic labels to their upper-level ontology topics by first retrieving candidate ontology topics and then manually identifying those that actually match. The matched ontology topics are then propagated to the corpora topic labels.

Using the 748 normalized topic labels as queries, we retrieve for each one of them 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 [138] index of the concatenated documents for each topic. BM25 is a modified version of the retrieval model TF-IDF that is widely used [42]. This enables us to narrow down the mapping of a normalized topic label to a manageable size. Except for a handful of cases, 50 candidate ontology topics were retrieved for each normalized topic label.

In an annotation task, we narrowed down the candidate ontology topics to those that are actually upper-topics or synonyms of a given normalized topic label—which thus indicates that all arguments in the corpus with that topic label are about

TABLE 4.7: Statistics of the mapped topic labels for each level of the three ontologies: the World Economic Forum (WEF), Wikipedia, and Debatepedia. For each level, we list the count of the topic labels that are mapped to each level, the count of all topics that are covered in each ontology level, as well as the min, mean, and max count of ontology topics that are mapped per normalized topic label.

Ontology level		**Ontology topics**			
	Topic labels	**All**	**Min**	**Mean**	**Max**
WEF Level 1	1,239	87	1	1.48	5
WEF Level 2	355	77	1	1.32	4
Wikipedia Level 1	1,539	14	1	1.24	3
Wikipedia Level 2	1,453	285	1	1.76	16
Debatepedia	2,002	87	1	2.80	10

the ontology topic. Three annotators were recruited for the task, where each annotator labeled the topics of one ontology. For each normalized topic label, an annotator labeled a candidate ontology topic for whether it is a synonym or an upper-topic to the normalized topic label. Each annotator labeled the candidate ontology topics separately for each level in the topic ontology. The annotators were allowed to map multiple ontology topics to a normalized topic label. For example, the normalized topic label “plastic bottles” is mapped to “pollution” and “recycling” in Wikipedia Level 2. To avoid ambiguity, we presented both the ontology topics and the topic labels with a topic description. We retrieved the topic descriptions from the topic ontology for the ontology topics and the first paragraph of the Wikipedia article for the normalized topic labels. The ontology topics for the normalized topic labels were then propagated to the corresponding topic labels in the argument corpora.

Analysis of Topic Coverage

Table 4.7 shows general statistics of this mapping of topic labels to ontology topics. Most of the topic labels (2,002 out of 2,117) are mapped to at least one Debatepedia topic, while only 355 labels are mapped to WEF Level 2 topics. For Wikipedia Level 2, only 285 out of the 748 topics are actually covered by argument corpora. Already this first analysis suggests that existing argument corpora typically cover a small subset of possible argumentative topics that people are trained to debate. For those topic labels that can be mapped are mapped on average to 2.8 topics in Debatepedia, to 1.24 topics in Wikipedia Level 1, and to 1.48 topics in WEF Level 1. As discussed in Subsection 4.2.2, topics in Debatepedia focus on the Western culture and are easily accessible, whereas topics in the WEF require deeper domain knowledge and have more global relevance. The high coverage of Debatepedia’s topics indicates that the studied argument corpora focus on common topics that are easily approachable, while global issues or those that need domain knowledge lack

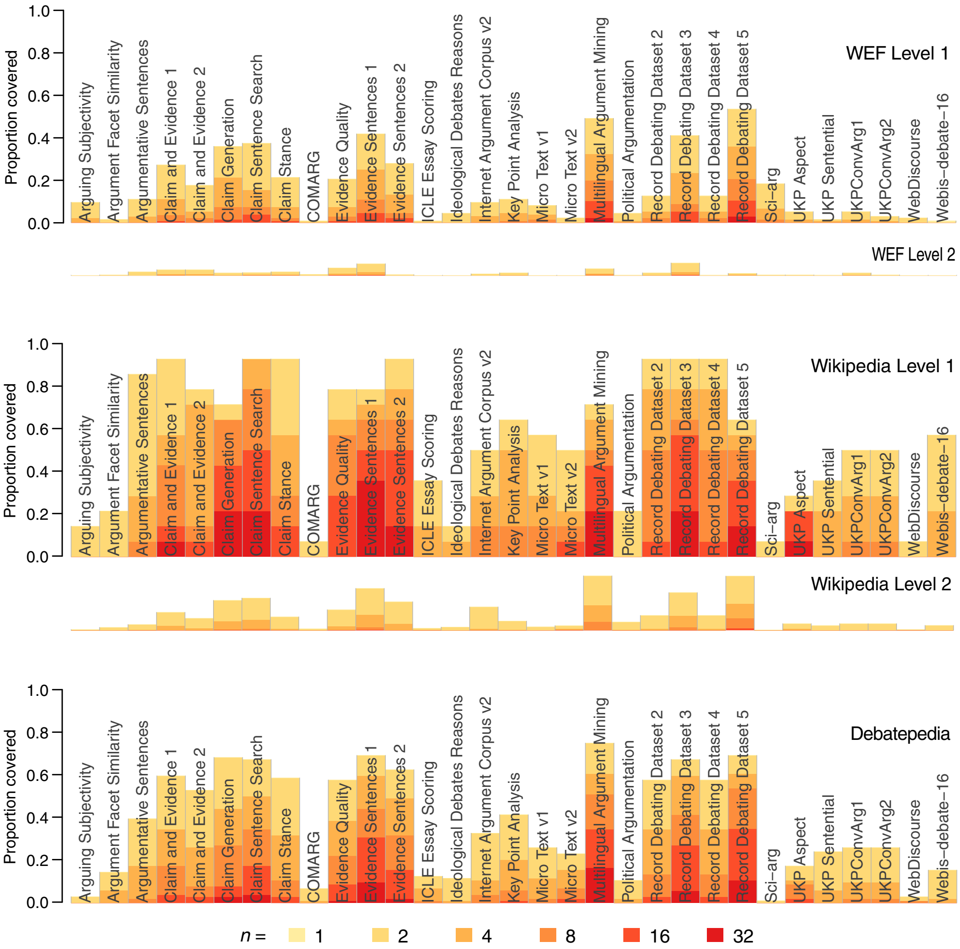


FIGURE 4.4: Proportion of ontology topics covered by at least n corpus topics (per ontology level and per corpus).

coverage.

For a more fine-grained analysis, Figure 4.4 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 4.5 shows how the set of all 2,117 corpus topics distribute over the top matching topics in Debatepedia, Wikipedia, and the WEF. The distribution is significantly skewed: while the top ten topics in Debatepdia are matched by 340 to 150 topic labels, the top ten topics in WEF Level 1 are matched by 125 to 50

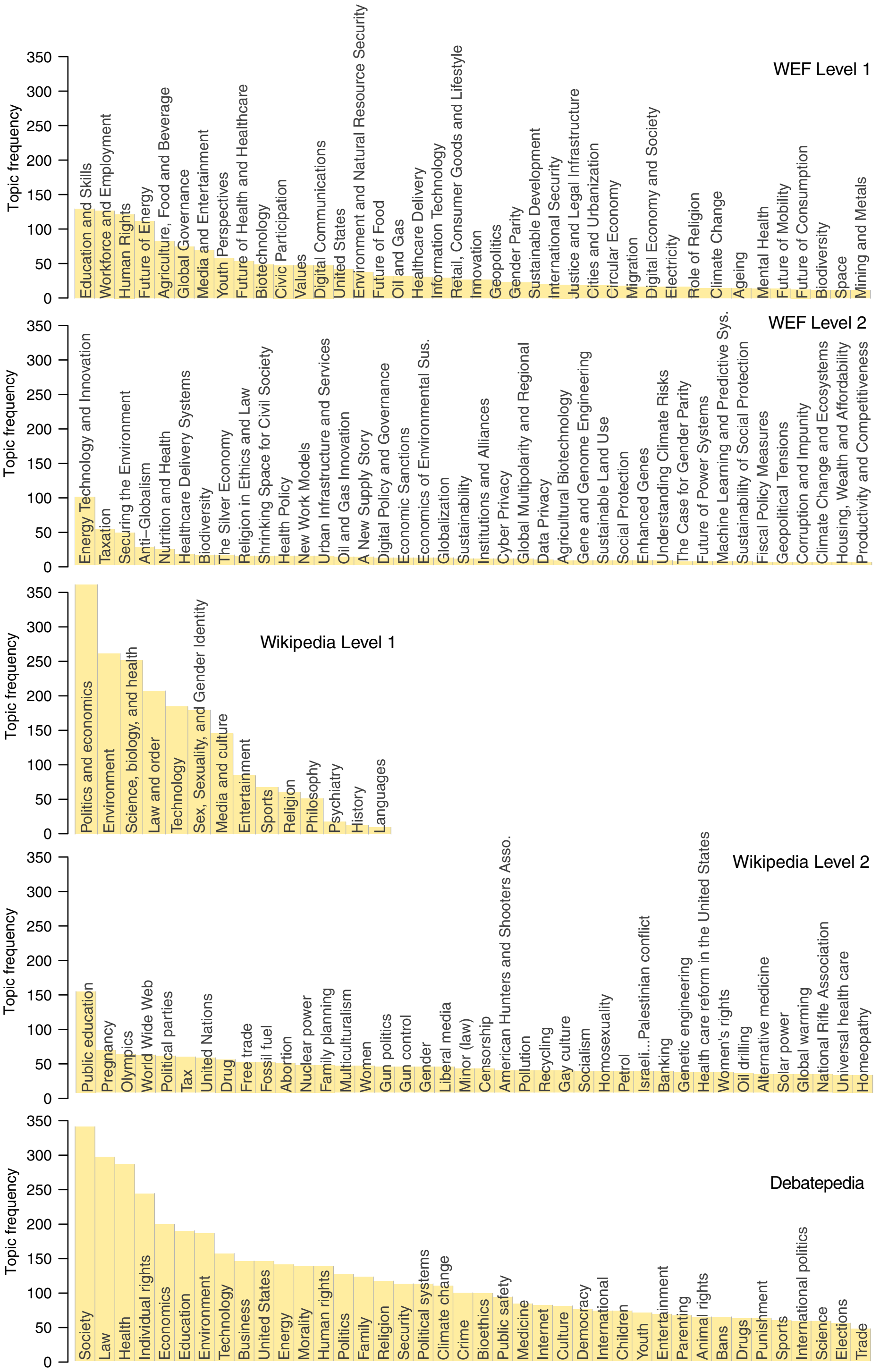


FIGURE 4.5: Distribution of 2,117 corpus topics over the top matching topics in an ontology (all corpora).

topic labels. The comparison between the three ontologies supports our previous finding that argument corpora cover easily accessible topics, especially "education and skills", "society", "politics and economics", "workforce and employment", "law", and "environment".

4.2.4 Corpus Unit Topic Categorization

The previous analysis on argument corpora is done on those corpora which contain topic labels. About a third of the argument corpora are thus excluded from that analysis. As a step toward assessing their topic coverage, we map the ontology topics for a unit (cf. Table 4.4) 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 crawled, concatenated, and used as the topic's representation. Though the documents associated with a topic are not necessarily argumentative, they can be expected to cover the salient aspects of the topic. To retrieve topics for a corpus unit, we implement and evaluate the following approaches:

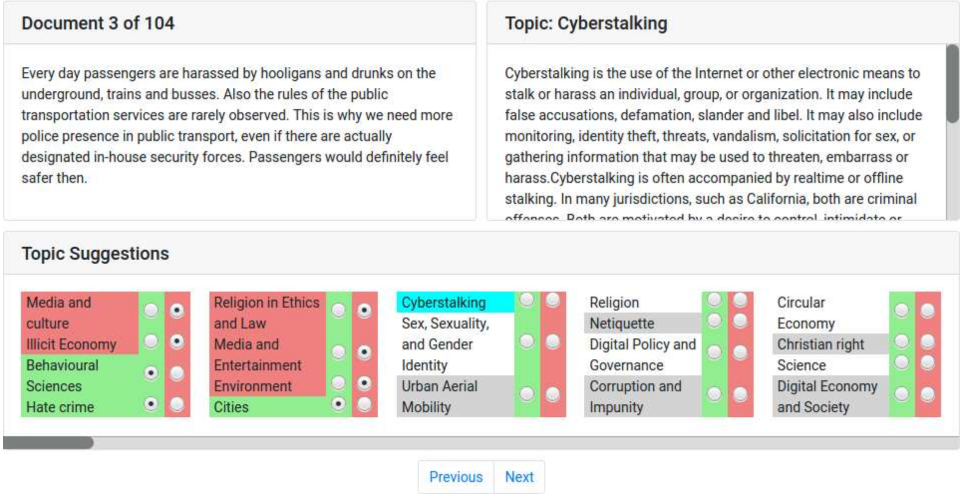
Semantic Interpretation (SI) This approach computes the semantic similarity of a unit and a topic as follows: it uses the cosine similarity of the TF-IDF vectors for the unit and the concatenated topic-related documents. This corresponds to the semantic interpretation step that is at the core of the well-known ESA model [59].

SI with Text Embeddings (T2V-SI) In Text2vec-SI, the similarity of topics and corpus units is calculated using text embedding vectors. We follow the common approach to generate text embeddings, which is to take the dimension-wise average of the word embeddings for all tokens in the text. We compare four embeddings:[6] the context-free GloVe embeddings [123] and three context-sensitive embeddings, namely ELMo [127], BERT [46], and the character-based flair (the news-forward-fast model; [8]).

Baselines Two baselines put our results into context: the random baseline classifies a unit per the prior probability of each ontology topic, whereas the direct match baseline does the same if the topic name appears in the unit text (ignoring case).

In order to assess the effectiveness of the approaches and baselines outlined previously, 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 anno-

[6] We use the default settings of the flair library version 4.5 [8] for all embeddings and truncate sentences at 200 tokens due to the maximum length for BERT. For efficiency, we limited the embeddings to 10,000 randomly sampled sentences for the topics that had more sentences associated with them.



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

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 4.6: The annotation interface for the pooled topic judgments. A corpus unit is presented on the left side, together with the most similar topics in all levels at the bottom. On the right side, the selected topic description is shown.

tated 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 only those topics which are among the five topics with the highest similarity to this unit according to at least one of the approaches (excluding the random baseline, which has been calculated from the results of this annotation). Figure 4.6 shows the employed assessment interface with 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 (relevant) and red (not relevant). For each unit, the topics from all levels were available in the annotation interface for annotation.

To reduce biases, both the units and the topics were shown in a different and random order to each assessor. Each topic was provided with a topic description from the ontology so that topics that are unfamiliar to the annotator can be understood. 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.

Based on the similarity scores of the approach, we derive Boolean labels that indicate whether a unit from the sampled units 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 policy with which the approach achieved the highest F1-score on the pooled judgments.

Table 4.8 shows the main results of this evaluation. The random baseline's
]

TABLE 4.8: Performance of the semantic interpretation (SI) and Text2vec-SI (T2V-SI) approaches and baselines in human evaluation for each topic ontology level in terms of precision (Pre), recall (Rec), and F1-score (F1) for the "about" label. For methods other than the baselines the table shows the values for the policy (threshold θ or rank k) that leads to the highest F1-score.

	Level 1				Level 2			
Approach	Policy	Prec	Rec	F1	Policy	Prec	Rec	F1
				WEF				
Random		0.02	0.02	0.02		0.01	0.01	0.01
Direct match		0.38	0.23	0.29		0.59	0.11	0.19
SI	$k = 12$	0.22	0.75	**0.34**	$k = 33$	0.21	0.70	**0.33**
T2V-SI$_{GloVe}$	$\theta = 0.98$	0.14	0.61	0.22	$\theta = 0.98$	0.12	0.66	0.21
T2V-SI$_{flair}$	$\theta = 0.96$	0.15	0.38	0.22	$\theta = 0.98$	0.24	0.18	0.21
T2V-SI$_{ELMo}$	$k = 4$	0.25	0.44	0.32	$k = 42$	0.13	0.68	0.23
T2V-SI$_{BERT}$	$k = 7$	0.19	0.53	0.28	$\theta = 0.93$	0.15	0.49	0.23
				Wikipedia				
Random		0.11	0.11	0.11		0.01	0.01	0.01
Direct match		0.12	0.04	0.06		0.47	0.34	0.40
SI	$k = 3$	0.32	0.65	0.43	$\theta = 0.05$	0.45	0.64	**0.59**
T2V-SI$_{GloVe}$	$k = 5$	0.19	0.66	0.30	$k = 705$	0.14	1.00	0.25
T2V-SI$_{flair}$	$k = 6$	0.17	0.68	0.27	$k = 692$	0.14	0.99	0.25
T2V-SI$_{ELMo}$	$k = 2$	0.39	0.53	0.45	$\theta = 0.76$	0.25	0.45	0.32
T2V-SI$_{BERT}$	$k = 2$	0.41	0.55	**0.47**	$\theta = 0.89$	0.22	0.52	0.31
				Debatepedia				
Random		0.07	0.07	0.07				
Direct match		0.47	0.34	0.40				
SI	$\theta = 0.02$	0.52	0.61	**0.56**				
T2V-SI$_{GloVe}$	$\theta = 0.98$	0.32	0.71	0.44				
T2V-SI$_{flair}$	$\theta = 0.93$	0.30	0.71	0.42				
T2V-SI$_{ELMo}$	$k = 13$	0.43	0.71	0.54				
T2V-SI$_{BERT}$	$k = 23$	0.36	0.80	0.50				

performance highlights the inherent difficulty of the task. The direct match 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 both baselines for all ontologies in terms of F1-score. The performance of the Text2vec approaches varies depending on the used embeddings, with ELMo and BERT being the most effective. The Text2vec approaches using BERT and ELMo outperform the baselines and the semantic interpretation approach on the most abstract topics (Wikipedia Level 1). On the second ontology levels, however, Text2vec approaches are subpar to the semantic interpretation approach and even fail to outperform the direct match baseline on Wikipedia Level 2.

The F1-score of the best approach (semantic interpretation) ranges depending on the ontology level from 0.33 to 0.59, leaving much room for improvement. Still, the classification performance should be taken in relation to the high number of topics, which ranges from 14 to 822 topics. A more effective approach to automatically identify the topics of a corpus unit can utilize the structure of the topic ontology using hierarchical classification. Hierarchical classifiers can classify the topic of documents starting from the upper level and then consider only its subtopics for classification in the lower levels.

Developing generalizable argument mining approaches relies on a careful sampling of the topics covered by the argument corpus on which the approach is developed. The three topic ontologies introduced in this section provide a way to assess the topic coverage of an argument corpus. In addition to using corpus construction standards, fostering the generalizability of an argument mining approach to new topics depends on how argument mining tasks are formulated. In the following, we introduce same side stance classification, which aims at making the task of identifying the stance of an argument less dependent on the topic.

4.3 Same Side Stance Classification

Identifying (i.e., classifying) the stance of an argument towards a particular topic is a fundamental task in computational argumentation and argument mining. The stance of an argument as considered here is a two-valued function: it can either be pro a topic (meaning, “yes, I agree”), or con the topic (“no, I do not agree”). Here we propose a related though simpler task, which we call *same side stance classification*. Same side stance classification deals with the problem of classifying two arguments as to whether they (a) share the same stance or (b) have a different stance towards the topic in question.

As an example, consider the following two arguments on the topic “gay marriage”, which obviously are on the same side.

Argument 1. Marriage is a commitment to love and care for your spouse till death. This is what is heard in all wedding vows. Gays can clearly qualify for marriage according to these vows, and any definition of marriage is deduced from these vows.

Argument 2. Gay marriage should be legalized since denying some people the option to marry is discriminatory and creates a second class of citizens.

Argument 3 below, however, is neither on the side of Argument 1 nor on the side of Argument 2.

Argument 3. Marriage is the institution that forms and upholds society. Its values and symbols are related to procreation. To change the definition of marriage to include same-sex couples would destroy its function because it could no longer represent the inherently procreative relationship of opposite-sex pair-bonding.

Same side stance classification (SSSC) is simpler than the "classical" stance classification problem, or at most equally complex: solving the latter implies solving the former as well.

Aside from the difference in problem complexity a second aspect renders same side stance classification a relevant task of its own right: Stance classification, by definition, requires knowledge about the topic that an argument is meant to address, i.e., stance classifiers must be trained for a particular topic and hence cannot be reliably applied to other (i.e., *across*) topics. In contrast, a same side stance classifier does not necessarily need to distinguish between topic-specific pro- and con-vocabulary; "merely" the argument similarity *within* a stance needs to be assessed. Consequently, same side stance classification is likely to be solvable independently of a topic or a domain—so to speak, in a *topic-agnostic* fashion. Since topic agnosticity is a big step towards application robustness and flexibility, we believe that the development of technologies that tackle this task has game-changing potential.

By presenting the SSSC task as a shared task, we evaluate approaches from eight German universities and IBM research for SSSC in two experiments: within a single topic and across two topics. The experiments are based on an argument dataset that is sampled from args.me corpus and covers two topics "gay marriage" and "abortion". After introducing the dataset, we present the approaches of the participants and compare their effectiveness in both experiments. Afterward, we conduct an error analysis to spot hard cases and easy cases that are faced by the approaches. At the end of the section, we conduct a manual inspection analysis of the task data, bringing to light its limitations and proposing several suggestions to enhance it.

TABLE 4.9: Number of argument pairs in the training set and the test set of the within-topic experiment.

	Training			**Test**		
Class	**Gay marriage**	**Abortion**	**Σ**	**Gay marriage**	**Abortion**	**Σ**
Sameside	13,277	20,834	34,111	63	63	126
Diffside	9,786	20,006	29,792	63	63	126
Σ	23,063	40,840	63,903	126	126	252

4.3.1 Dataset

Because of its size and the balanced stance distribution, the args.me corpus provides a rich source for our experiments. At the time of the shared task the corpus consisted of 387,606 arguments that were collected from 59,637 debates; a detailed description can be found in [6].[7]

An argument in args.me is modeled as a conclusion along with a set of supporting premises. In addition, each premise is labeled with a stance, indicating whether it is pro or con the conclusion. The stances originate from the debates where the arguments are used in. Debates can be started from different viewpoints, for instance, a debate may discuss the viewpoint “abortion should be legalized” while another may discuss “Abortion should be banned.”). Therefore, the stance of an argument has to be interpreted in relation to the arguments in the same debate. During the acquisition process of the data for the shared task we followed this constraint by ensuring that the arguments of an argument pair stem always from the same debate.

The count of debates that treat “abortion” and “gay marriage” is 1,567 and 712, respectively. We filtered out those arguments whose premises are shorter than four words since they are often meta statements such as “I win” or “I accept”. As a result, we kept 9,426 arguments on abortion and 4,480 arguments on gay marriage for the task.

4.3.2 Experiments

Starting from the arguments in a debate, we generated all possible argument pairs. An argument pair was labeled as *Sameside* if both arguments are either pro or con the viewpoint of the debate. Otherwise, the pair is labeled as *Diffside*. Pairs with identical arguments were removed.

[7] The entire args.me corpus can be accessed here: `https://webis.de/data.html#args-me`

TABLE 4.10: Number of argument pairs in the training and test set of the cross-topic experiment.

Class	Training: Abortion	Test: Gay marriage
Sameside	31,195	3,028
Diffside	29,853	3,028
Σ	61,048	6,056

Within-topic Experiment The within-topic experiment covers both topics in its training and test sets. The training set contains 67% of the argument pairs of one topic, which were randomly chosen. The test set was formed from the remaining 33% for the respective topic. Among others, it was ensured that a label for an argument pair in the test set can not be transitively deduced.[8] Note in this regard that the "same side" relation forms an equivalence relation. See Table 4.9 for the within-topic dataset statistics.

Cross-topic Experiment The cross-topic experiment provides a different topic for training from the one for testing. In particular, the training set contains argument pairs from the "abortion" debates only, while the test set contains argument pairs from "gay marriage" debates only. Sameside pairs and diffside pairs are balanced. See Table 4.10 for the cross-topic dataset statistics.

4.3.3 Submission

Overall, nine teams participated in the first shared task on same side stance classification. This section provides a brief overview of the approaches that the teams submitted, along with their results.

Düsseldorf University The approach followed by Düsseldorf University relies on a Siamese network trained to predict the similarity of two arguments on top of a small BERT [46]. As the maximum token length for BERT is 512 tokens, a relevance selection component to rank sentences by relevance is integrated, cutting the ranked input at 512 tokens. The approach achieved an accuracy of 60% on the within-topic task and 66% across topics.

IBM Research The approach submitted by IBM is based on a small vanilla BERT model and has been first fine-tuned to perform standard binary pro/con stance classification on data extracted from the IBM Debater project. On top of this model, another model is initialized and fine-tuned on the same side classification task.

[8] With transitive deduction we mean: $\mathit{SameSide}(A_1, A_2) \quad \wedge \quad \mathit{SameSide}(A_3, A_2) \quad \vdash \mathit{SameSide}(A_1, A_3)$

Table 4.11: The results of the submissions for the within-topic experiment and the cross-topic experiment in terms of precision, recall, and accuracy. For both Trier University† and MLU Halle‡, the best and the worst result are reported since they submitted multiple approaches.

	Within-topic			Cross-topic		
Team	**Precision**	**Recall**	**Accuracy**	**Precision**	**Recall**	**Accuracy**
Trier University†	**0.85**	0.66	**0.77**	**0.73**	**0.72**	**0.73**
Leipzig University	0.79	0.73	**0.77**	0.72	**0.72**	0.72
IBM Research	0.69	0.59	0.66	0.62	0.49	0.60
TU Darmstadt	0.68	0.52	0.64	0.64	0.59	0.63
Düsseldorf University	0.70	0.33	0.60	0.72	0.53	0.66
Trier University†	0.65	0.24	0.56	0.70	0.11	0.53
LMU	0.53	**1.00**	0.55	0.67	0.53	0.63
MLU Halle‡	0.53	0.57	0.54	0.50	0.57	0.50
Paderborn University	0.59	0.19	0.53	0.60	0.38	0.56
University of Potsdam	0.51	0.58	0.51	0.51	0.52	0.51
MLU Halle‡	0.50	0.11	0.50	0.46	0.00	0.50

The approach obtained results inverse to the ones of Düsseldorf University: 66% accuracy in the within-topic setting and 60% in the cross-topic setting.

Leipzig University The approach submitted by Leipzig University uses a pre-trained BERT model that is fine-tuned on the same side stance classification task. In addition, a binary classification layer with one output and cross-entropy loss function is used instead of a multilabel classification layer. To embed an argument, the first 254 tokens of an argument are fed through the BERT model. Then, the last 254 tokens of an argument are embedded. The concatenation of both embeddings is fed into the classification layer. The approach achieved an accuracy of 77% in the within-topic setting and 72% in the cross-topic setting.

LMU The approach submitted by the Ludwig Maximilian University (LMU) relies on a vanilla pre-trained BERT base model that is fine-tuned to the shared task. The data is organized in a graph with one graph per topic. Nodes represent arguments, and edges are labeled with the confidence that the associated arguments agree with each other. This graph-based approach has the benefit that more training data can be generated by a transitive closure. Its accuracy was 55% in the within-topic setting and 63% in the cross-topic setting.

MLU Halle The approach submitted by the Martin-Luther-University (MLU) of Halle-Wittenberg consists of three approaches. The first approach uses a tree-based learning algorithm as classifier using standard bag-of-words features. The second is a rule-based approach that reduces the task to sentiment classification relying on rules defined over lists of words with their polarity taken from a sentiment lexicon. The third is a re-implementation of the stance classification approach of Bar-Haim et al. [19]. The best approach achieves an accuracy of 54% on the within-topic setting and 50% in the cross-topic setting.

Paderborn University The approach used by Paderborn University relies on a Siamese Neural Network to map arguments to a new space where arguments with the same stance are closer to each other, and other arguments are less close. Arguments are represented by the contextual word embeddings provided by the flair library [8]. A final sigmoid activation function produces the output used for same side stance classification. The approach achieved an accuracy of 53% within topics and 56% across topics.

Trier University The approach submitted by Trier University relies on a pre-trained BERT base model fine-tuned to the shared task. It was submitted with different configurations. The best yielded an accuracy of 77% in the within-topic setting and 73% on the cross-topic setting, the worst 56% and 53%, respectively.

TU Darmstadt The approach followed by the Technical University of Darmstadt relies on a multi-task deep network that is based of the pre-trained large BERT model. The network is trained on a number of pro/con stance classification datasets in addition to the shared task dataset. The approach achieved an accuracy of 64% in the within-topic setting and 63% in the cross-topic setting.

University of Potsdam The approach submitted by the University of Potsdam relies on Bi-LSTM to encode the arguments. The embeddings of both arguments are concatenated, multiplied in an element-wise fashion, subtracted, and fed into a two-layer MLP as a classification layer. The approach achieved 51% accuracy both within and across topics.

Discussion

The results of the shared task license a number of interesting conclusions. First, the results have validated our hypothesis that a topic-agnostic approach to same side stance classification is feasible. This is clearly conveyed by the fact that the within-topic and the cross-topic setting seem to be of similar complexity. Also, the differences in accuracy on both tasks are less than 5–6% points, additionally corroborating the hypothesis.

A second conclusion is that the effectiveness of most approaches clearly improves over a random baseline, showing that the task is generally feasible. At the same time, however, the results show that there is potential for improvement.

4.3.4 Error Analysis

In this section, we present the outcomes of manually analyzing the predictions of the nine approaches submitted to the shared task. We examine the argument pairs which are classified correctly (or wrongly) by most of the approaches. A careful review of these pairs reveals some easy and hard cases for the same side stance classification. In the following, we discuss these cases in detail.

Easy Cases

In total, we found 1,234 pairs in which all the submitted approaches classified correctly, 1,215 in the cross-topic experiment, and 19 pairs in the within-topic. From these pairs, we determined four cases where classifying the same side stance is doable computationally (i.e., easy cases):

1. The stance toward the same topic is expressed explicitly in the two arguments:

   Argument 1. ...because I don't believe in gay marriage ...

   Argument 2. ...I want to first off point out that I am against gay marriage personally ...

2. The two arguments include contradicting statements:

   Argument 1. ...marriage is not a recognition of love and compassion ...

   Argument 2. Marriage is about love. ...

3. An argument questions a certain statement in the other argument:

   Argument 1. People should be allowed to make their own choices in life without having their human rights taken away.

   Argument 2. I would like to know how people making their own choices has their rights taken away in the first place. Give me something to argue about!

4. An argument quotes a certain statement in the other argument:

Argument 1. I also gave references stating that in the bible homosexuality isn't even accepted.

Argument 2. *"I also gave references stating that in the bible homosexuality isn't even accepted"* oops - sorry - the bible isn't admissible as a source of law in the us.

Hard Cases

In the test dataset, 126 argument pairs were difficult to be classified by the approaches (125 in the cross-topic experiment). Two cases were noticeable in these pairs:

1. Further knowledge about the discussed topic is needed to resolve the stance:

Argument 1. Gay marriage violates religious freedoms.

Argument 2. Gay marriage is a negligible change to the institution of marriage.

2. The two arguments agree on one aspect related to the topic but disagree on other aspects:

Argument 1. Marriage is a euphemism for using the government to enforce a relationship. There's no problem with gays getting married, but they shouldn't marry with government involvement.

Argument 2. I say we let the gays get married. It's not like it affects anyone but them anyway.

4.3.5 Data Quality

The shared task dataset is derived from args.me corpus [6]. This corpus incorporates five different debate platforms: four comprise arguments in a monological form, while one embraces arguments within dialogues (aka debates). Because the latter is the largest platform that contributes the most to the args.me corpus with more than 182,198 arguments (63%), it largely dominates the shared tasks datasets.

Deriving arguments from dialogues, however, requires extensive text normalization, including removing meta-dialogue and meta-user information, filtering low-quality texts that contain abusive language or spam, and de-contextualizing arguments.

This preprocessing step was not performed for the shared task datasets, which led to several invalid argument instances. Overall, we found two main problematic cases:

1. The argument addresses solely a debate meta-information:

   > **Argument .** This round is for acceptance only. The rest will be for argumentation.

   > **Argument .** My opponent had forfeited the round, so my arguments stand unchallenged.

2. The argument contains ad hominom attack:

   > **Argument .** Like I said I didnt copy crap! and if you are going to accuse me for something I didn't do, then I wish to never have another debate with you again.

Given that these cases frequently occur in the shared task datasets, we suggest the following improvements:

- Using only monological sources of arguments, as dialogues need the preprocessing step we mentioned above.
- Conducting manual annotation or validation of the argument pairs, especially for those which are put in the test datasets.

4.4 Summary

Enabling argument retrieval systems to respond to argumentative queries requires a proper source of argumentative content. While the web offers the broadest coverage of argumentative content, it is characterized by different genres. In Section 4.1, we compared different approaches for the task of argument unit segmentation and assessed their generalizability across genres. We cast this task as token-level sequence labeling and compared different token-level features and sequence-to-sequence models to perform the task. We found that semantic and structural features are the best for detecting the boundaries of argumentative units across genres. We also found that a sequence-to-sequence model that captures a wider context tends to perform better within and across genres. Still, the results show that the employed linguistic features and machine learning models do not generalize well across genres.

A realistic application of an argument mining approach to provide arguments to a retrieval system is challenged by its ability to generalize over topic. Guaranteeing the generalizability of argument mining approaches to new topics requires careful sampling of topics while constructing argument corpora. To this end, we introduced three topic ontologies that are tailored for argumentation and created by domain experts. Using the topic ontologies, we analyzed the topic coverage and distribution of 31 argument corpora, which are all existing argument corpora that are provided with topic labels. The analysis showed that the topic distribution of these argument corpora is skewed and concentrated around a small set of topics.

An analysis of the topic coverage of argument corpora with no topic labels is bound to develop automatic approaches that map a corpus unit to its corresponding ontology topics. Toward this goal, we developed several approaches that take a corpus unit and a topic as input and return a score that quantifies how likely can the corpus unit be used as an argument about the topic. We manually evaluated the pooled output of the approaches and two baselines for a sample of 104 corpus units. The classification performance of the best approach (semantic interpretation) on the three topic ontologies ranges from 0.33 to 0.59 F1-score, leaving much space for improvement. Future research can utilize hierarchical classifiers to reduce the high topic count in the lower levels of a topic ontology.

Apart from adopting topic selection standards, generalizability to new topics should be guaranteed while formulating argument mining tasks. To tackle the topic-dependence of stance classification, we introduced a new formulation of the task that takes a pair of arguments as input and returns whether they are on the same or opposite side. We solicited nine approaches from eight German universities and IBM Research for the task. To assess how well the approaches generalize over topic, we designed cross-topic and within-topic experiments using a dataset that covers two topics ("abortion" and "gay marriage"). The best approaches in both experiments used BERT [46], but differently handled the long length of the arguments in the dataset in comparison to the length allowed by BERT. The results of the best approaches show a very close performance between the in-topic and cross-topic experiments (about 5% difference in terms of accuracy). This supports our hypothesis that a topic agnostic approach for stance classification is feasible. An analysis of the easy cases and hard cases classified by the best approaches shows that missing knowledge in argument pairs and partial agreement/disagreement between them are the main challenges in the task.

Chapter 5

Identification of Argument Frames

Persuading an audience with a stance requires a careful constellation of arguments that are tailored to the target audience. Selecting and phrasing arguments in a way that emphasizes certain aspects and hides others is known as framing. Argument retrieval systems retrieve a list of arguments as pro and con and rank them according to their relevance and quality. Delivering arguments with their frames allows a user to locate arguments that appeal to the target audience that they are addressing. This section first introduces an approach to identify the frames of an argument and then proposes a visual interface to present and explore the retrieved arguments by their aspects.

5.1 Frame Identification

While producing an argumentative text (e.g., a persuasive speech), the author has to choose from numerous arguments that exist for a given topic. By choosing among the available arguments on a topic, the author frames the topic by emphasizing a specific aspect while concealing others. For instance, the following arguments target different topics but concentrate on the same frame, namely, the "economic" aspect.

Argument 1 *"I support the legalization of marijuana since it can be taxed for revenue gain."*
Topic: Marijuana

Argument 2 *"Legalizing prostitution would increase government revenue. A tax on the fee charged by a prostitute and the imposition of income tax on the earnings of prostitutes would generate revenue."*
Topic: Prostitution

Framing is a decisive step in the construction of an argument, which determines its persuasive effect on a given audience [49]. To achieve persuasion, an author of an argumentative text should choose frames that resonate with the target audience. As a simple example, an argument appealing to Christianity might not be acceptable to an atheist. Knowing the arguments for a topic along with their frames enables authors to choose those arguments that best address their audience.

This section introduces an unsupervised approach to identify frames in arguments that are assumed to cover a variety of topics. The approach is based on a formal view that defines a frame to be a set of arguments that share an aspect. More specifically, a frame F is a subset of a set of arguments A, $F \subseteq A$. Likewise, a set of frames, $\{F_1, \ldots, F_k\}$ covers a set of arguments iff. $A \subseteq \bigcup_j^k F_j$. Starting from a set of arguments, our approach first clusters them into topics, removes topical features from the arguments, and then clusters the arguments into frames. To evaluate the approach, we introduce a dataset of 12,326 arguments, which are labeled with 330 *generic frames* (frames that are used in multiple topics) and 1,293 *topic-specific frames* (those that are used only in one topic). We apply the approach to all arguments in our dataset and evaluate the returned frames against the ground-truth frames of the arguments. At the end of the section, we analyze the errors made by the approach in the experiments.

The contributions provided in this section cover:

- A formal view of frames in argumentation.
- An unsupervised approach to identifying frames in a set of argumentative texts.
- An argument framing dataset with 465 topics, 1,623 frames, and 12,326 arguments.

We freely provide the complete dataset to the research community.[1]

5.1.1 Data

Debate portals are websites where people debate or collect arguments for or against controversial topics. Some debate portals are dialogical, such as `debate.org`, allowing two opponents to debate one topic in rounds. Other debate portals are wiki-like (e.g., Debatepedia), where arguments are listed according to their stance on the topic. Debate portals keep a canonical structure of the arguments considered for each topic (usually a conclusion and a premise). The structure and the high quality of argumentation offered by debate portals have made them a suitable resource for research on computational argumentation [9, 32, 173].

Argument Frames from Debatepedia

For the given work, we crawled all arguments from Debatepedia in order to construct a dataset for the evaluation of frame identification. Debatepedia organizes a

[1] `https://webis.de/data/webis-argument-framing-19.html` or `https://doi.org/10.5281/zenodo.3373355`

TABLE 5.1: Counts of topics, frames, merged frames, and arguments in the Webis-Argument-Framing-19 dataset.

Topics	Frames	Merged frames	Arguments
465	1,645	1,623	12,326

debate into sets of arguments that address a topical aspect of the debate. A label that describes the topical aspect is attached to some of the sets, such as "economics". An argument on Debatepedia is listed as a conclusion on the topic along with a premise that supports it.

Arguments that are not labeled might introduce noise to the dataset, since the true knowledge regarding their frames is unavailable. To exclude possible noise in the planned experiments, we filtered out all arguments without labels (about 1,800). Next, we analyzed the extracted labels and found that some labels have a similar meaning but are worded differently. In particular, we noticed the presence of the following cases:

1. Labels with hierarchical relations, such as "business" and "US business".
2. Opposite labels, such as "health" and "unhealthy", or, "protecting smokers" and "protecting non-smokers".
3. Labels that are equal when being lemmatized, such as "economics" and "economic", "democratizing" and "democratic", etc.

Labels with the same lemmas are likely to carry the same meaning, which is why we merged them into the same label. The count of such merged label pairs was 22, each containing 42 arguments on average. Since the labels in the first and second cases might constitute different frames in some contexts, we kept them as they are.

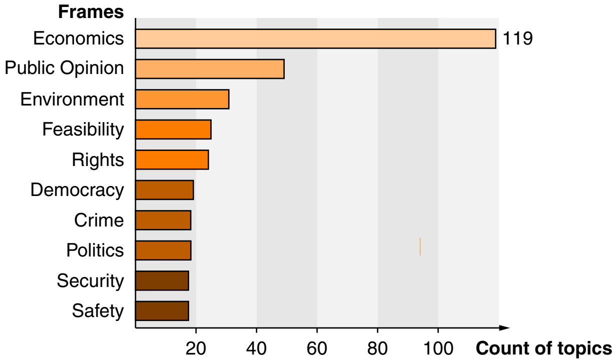


FIGURE 5.1: The number of topics in which each of the ten most frequent frame labels in our dataset occurs.

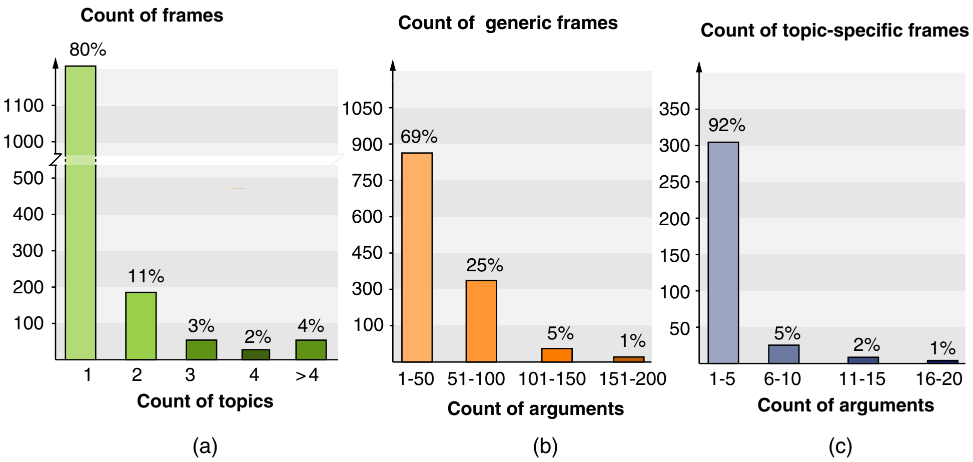


FIGURE 5.2: General statistics of frames from Debatepedia in the Webis-Argument-Framing-19 dataset. (a): Histogram of frames over the count of topics in which they are used. (b): Histogram of generic frames over the count of arguments they contain.
(c): Histogram of topic-specific frames over the count of arguments they contain.

Webis-Argument-Framing-19 Dataset

Table 5.1 shows general statistics of the final dataset after crawling and preprocessing, which we call *Webis-Argument-Framing-19*. As visualized in Figure 5.1, the ten most frequent labels in our dataset are: “economics”, “public opinion”, “environment”, “feasibility”, “rights”, “democracy”, “crime”, “politics”, “security”, and “safety”. These labels largely overlap with those introduced by Card et al. [35]; hence, we considered each set of arguments to be a frame.

The count of topics in which a frame occurs indicates whether a frame is generic or topic-specific. To distinguish between these two types of frames, we grouped all frame labels in our dataset according to how many topics they are used for. Figure 5.2 (a) shows a histogram of the frames in our dataset over the count of topics in which they are used. As depicted, 80% (1,293) of the frames are used in one topic and, hence, we labeled them as *topic-specific*. Frames that are used in more than one topic add up to 20% (330) frames and are labeled as *generic*. Generic frames in the dataset cover 7,052 arguments, while topic-specific frames cover 5,274 arguments. Figure 5.2 (b) and (c) show a histogram of generic and topic-specific frames over the count of arguments they contain, respectively. The histograms reveal that generic frames cover an order of magnitude more as many arguments as topic-specific frames.

5.1.2 Approach

In this section, we introduce our unsupervised approach to modeling frames formally. We assume frames to be exclusive and non-overlapping. Given a set of

TABLE 5.2: Notation of the symbols used in the approach

Symbol	Meaning
a	An argument
c	The conclusion of an argument
A	A set of arguments
$\bar{A}$	A set of arguments on the same topic
$\mathcal{A}$	A set of sets of arguments
F	A frame
v	A word
V	A vocabulary
E	A topic extraction model

arguments $A = \{a_1, a_2, \dots, a_n\}$, our goal is to find a set of frames that constitutes a cover of A. A cover of A is a set of sets $\{F_1, F_2, \dots, F_k\}$ whose union contains A, i.e., $A \subseteq \bigcup_j^k F_j$. Table 5.2 lists the symbols used in this section along with their meaning.

The main idea of our approach is to first remove topical features from arguments and then to cluster the arguments into frames. Following known topic modeling approaches, we represent the content of an argument a as a bag of words and propose two models to find topic-specific words. Both models utilize the frequency of the words in an argument and the argument's structure. The structure of a is represented by its conclusion c and its premise(s) p. Our approach includes three main steps:

(a) **Topic clustering.** Cluster the arguments in A into m topics $\mathcal{A} = \{\bar{A}_0, \bar{A}_1, \dots, \bar{A}_m\}$.

(b) **Topic removal.** Given the produced clusters, develop an extraction model E that extracts topical features from an argument a_i and its cluster. E is applied to each $\bar{A}_j \in \mathcal{A}$ to remove topic-specific features. As a result, we obtain "topic-free" arguments $a'_i = a_i - E(a, \bar{A}_j)$. We denote the set of all "topic-free" arguments with A' where $A' = \{a'_1, a'_2, \dots, a'_n\}$.

(c) **Frame clustering.** Cluster the arguments A' into k clusters, each respresenting one frame.

Figure 5.3 sketches the general idea of the three steps of our approach. We detail our concrete realization of each step in the following.

Topic Clustering

To cluster the given set of arguments into topics, we first map each argument into a vector space that represents its semantics. We use k-means [75] with Euclidean

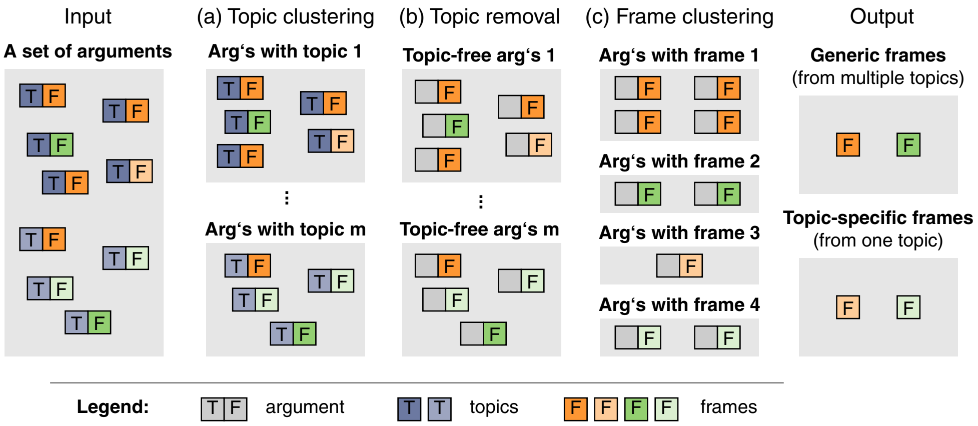


FIGURE 5.3: Sketch of the proposed unsupervised approach to argument frame identification. An argument is modeled as a topic and a frame. The input is a set of arguments. The output is a representation of two types of found frames: generic frames and topic-specific frames.

distance as clustering algorithm. For semantic spaces, we consider two alternatives: Term Frequency-Inverse Document Frequency (TF-IDF) and Latent Semantic Analysis (LSA).

TF-IDF TF-IDF defines a vector space whose dimensions are words in the dataset. An argument is mapped to this space according to the frequency of each of its words, normalized by the word's frequency in all considered arguments. TF-IDF is a sparse vector space since all words in a set of arguments are considered. To reduce sparsity, we construct a vocabulary V that comprises the 5,000 most frequent words in the arguments after stopword removal. Words that occur in more than half of the arguments are ignored as well. The main reason for reducing the vocabulary is to increase the computational efficiency of the approach.

LSA Latent Semantic Analysis [45] infers from a term-document frequency matrix a linear transformation that projects documents into a topic space. We construct two different semantic spaces using LSA. The first, simply called LSA, considers each argument to be a document. The second, *LSA Debate*, considers a whole debate to be a document. Since LSA Debate works on the debate level, it can better capture the topic context of an argument. The reason is that arguments capture the topic differently and may have few words in common. Using all arguments in a debate ensures a broader context of the topic. To compare both LSA models systematically, we use the same number of dimensions for both models: 1,000.

Topic Removal

The goal of this step is to remove topic-specific features in the topic clusters $\mathcal{A} = \{\bar{A}_0, \bar{A}_1, \ldots, \bar{A}_m\}$. To achieve this goal, we develop two models to extract topic-specific features, E_1^q and E_2. E_1^q utilizes the content of the arguments in one cluster, whereas E_2 utilizes the argument structure, i.e., conclusion and premise information.

E_1^q utilizes the term-frequency inverse document frequency measure TF-IDF for every word v in each cluster. We calculate idf as follows:

$$idf(v) = \frac{|\mathcal{A}|}{\big|\{\bar{A}_j \in \mathcal{A} : v \in \bar{A}_j\}\big|}$$

Then, $E_1^q(a)$ returns those words that best discriminate a specific topic based on a threshold q, which can be understood as the "aggressiveness" of the model, as follows:[2]

$$E_1^q(a, \bar{A}_j) = \{v \in \bar{A}_j : tf.idf(v) > q\}$$

E_2 utilizes the structure of an argument on a local level. The hypothesis here is that the conclusion c of an argument a contains more words that target the topic than its premise. Hence, we remove the conclusion in an argument. Formally:

$$E_2(a, \bar{A}_j) = \{v \mid v \in c\}$$

Frame Clustering

This step aims at grouping arguments that share a common aspect after removing topical features. For clustering, we use k-means again and experiment with different values of k. Below, we choose k based on an experiment that evaluates the output of the cluster against the ground-truth. We also use Euclidean distance to estimate the similarity between the arguments in the two semantic spaces.

5.1.3 Experiments

Based on the dataset we introduced in Section 5.1.1, we conduct experiments to evaluate and analyze our approach to modeling frames in argumentation. As discussed above, the approach consists of three steps: topic clustering, topic removal, and frame clustering. We evaluate the three steps and their interaction with each other in different experiments.

[2] In our experiments, we chose the threshold q empirically.

Topic Clustering Experiment

The goal of this experiment is to find the best method to group arguments into topics. The produced clusters for each semantic space are evaluated against the arguments' topics in the ground-truth dataset. An external measure is then used to evaluate the output of the clustering algorithm for each semantic space. In particular, we use Bcubed F1-score [17] to evaluate the effectiveness of our approach in modeling topics in the dataset. Bcubed F1-score rewards only the instance pairs that exist in the output of the clustering algorithm and in the ground-truth together in the same cluster. The reason for choosing Bcubed F1-score is that it is proven to satisfy desired constraints in the output of clustering algorithms [13].

Topic Removal Experiment

This experiment evaluates our models E_1^q and E_2 at removing topical features from the arguments in $\mathcal{A}$. The evaluation criterion here is the effectiveness drop of the topic clustering algorithm after removing the topical features in $\mathcal{A}$. We rerun the topic clustering algorithm with the same k after removing the output of both models E_1^q and E_2. To have a consistent comparison, we set k to the best count of topics we found in the previous experiment.

Frame Clustering Experiment

The last experiment evaluates clustering arguments into frames after topic removal. To test our hypothesis that topic removal benefits frame identification, we also cluster arguments in the same semantic space without topic removal. For both semantic spaces, we conduct three experiments: main experiment, generic experiment, and topic-specific experiment. In the topic-specific and generic experiment, we use the frames in our dataset that are labeled as topic-specific and generic frames separately. In the main experiment, we test our approach on the whole dataset without distinguishing the type of frames. The different experiments should show us the performance of our approach at identifying generic and topic-specific frames. Similar to topic clustering, we use Bcubed F1-score [17] to evaluate the frame clustering algorithm in the three experiments. Since our dataset contains 1,623 frames, we evaluate the output of the clustering algorithm for each $k \in \{100, 200, \ldots, 1,600\}$.

5.1.4 Results and Discussion

In the following, we report on the results of the three experiments explained above separately. In the end, we discuss the findings of the experiments and draw final conclusions on the performance of our approach at identifying frames.

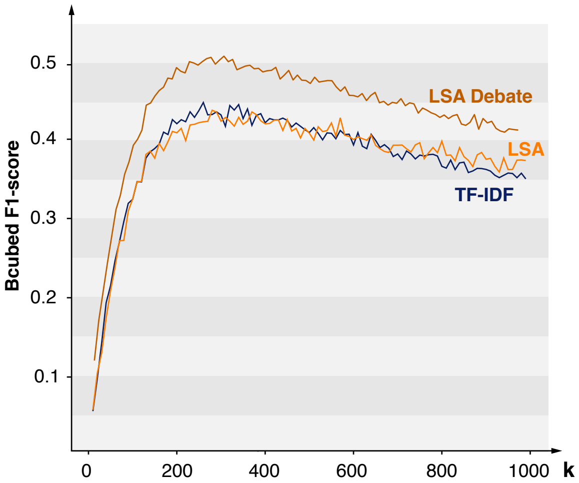


FIGURE 5.4: Bcubed F1-score of the topic clustering algorithms for the semantic spaces TF-IDF, LSA, and LSA Debate for each k.

TABLE 5.3: Bcubed F1-score of the topic clustering algorithm for each semantic space and the corresponding count of topics found.

Semantic space	Count of topics	F1-score
LSA Debate	310	**0.52**
TF-IDF	260	0.45
LSA	280	0.44

Topic Clustering

Figure 5.4 shows the effectiveness of topic clustering using the different semantic spaces. We visualize for each k the Bcubed F1-score of the clustering algorithms for the three semantic spaces. As shown, TF-IDF and LSA perform similarly for all k. The clustering algorithm performs better using the semantic space LSA Debate than LSA and TF-IDF. This shows the importance of considering the context of an argument for modeling their topics. All the three depicted plots, however, show a clear elbow between topic counts 200 and 400. Table 5.3 shows the highest corresponding F1-score and the count of topic clusters for each semantic space. The best topic clustering achieved by the algorithms comprises 310 clusters. Given its high effectiveness in modeling topics, we decided to proceed with the topic clusters produced by the LSA Debate in the next experiment.

Topic Removal

Table 5.4 shows the results of the topic removal experiment and frame clustering experiment. For both semantic spaces, the effectiveness of the topic clus-

TABLE 5.4: Best Bcubed F1-score, precision, and recall for the topic extraction models E_1^q, E_2, and without topic removal (baseline) in the generic, topic-specific, and main frame experiments together with the corresponding Bcubed F1-score (F1) in topic clustering in the semantic spaces TF-IDF and LSA.

Topic removal		**Frame clustering**								
Model	**Topic**	**Generic frames**			**Topic-specific frames**			**Frames**		
	F1	F1	P	R	F1	P	R	F1	P	R
					TF-IDF					
Baseline	**0.45**	0.19	0.25	0.15	**0.48**	0.53	0.44	0.26	0.27	0.25
$E_1^{0.005}$	0.42	**0.28**	0.26	0.30	0.45	0.50	0.40	**0.28**	0.24	0.33
E^2	0.17	0.26	0.25	0.28	0.45	0.48	0.42	0.27	0.25	0.29
					LSA					
Baseline	0.44	0.16	0.20	0.13	0.39	0.44	0.35	0.21	0.22	0.22
$E_1^{0.005}$	0.4	0.21	0.15	0.33	0.47	0.44	0.48	0.26	0.25	0.27
E^2	0.25	0.2	0.18	0.22	0.46	0.41	0.50	0.24	0.24	0.24

tering algorithm is reported after using the models E_1^q and E_2 to remove topic-specific words. To evaluate the topic extraction models, we re-list the effectiveness achieved by the topic clustering algorithm for both spaces. We show the results of E_1^q only for $q = 0.005$ since higher values of q showed similar or lower results in all experiments.

As shown, E_2 decreases the effectiveness of the topic clustering algorithm to about the half. The model $E_1^{0.005}$ achieves a smaller drop of 0.03-0.04 in the two semantic spaces. Despite its simplicity, E_2 is more effective at removing topic-specific features than $E_1^{0.005}$.

Frame Clustering

Table 5.4 shows the results of the frame clustering algorithm in the experiments: generic, topic-specific, and main. In each experiment, the clustering algorithm is run after using the two topic extraction models to remove topic-specific features and without applying them (baseline). In the main and the generic experiment, using the topic extraction models outperforms not using them in both semantic spaces. In the topic-specific experiment, our approach's effectiveness outperformed the baseline only in the LSA space. The comparison between the results in the generic and topic-specific experiments shows that identifying generic frames is harder. The reason can also be the small size of topic-specific frames in the ground-truth. Our approach, however, is only effective at identifying generic frames and fails at outperforming the baseline in the topic-specific experiments. A reason to justify this is that removing topic-specific features negatively affects identifying topic-specific frames.

To better analyze our approach, we plot the achieved Bcubed F1-score for each semantic space and each experiment for different values of k. Figure 5.5 (a, c, e) shows the effectiveness achieved in the three experiments: main, topic-specific, and generic in the TF-IDF space, respectively. As shown, both models $E_1^{0.005}$ and E_2 start to outperform the baseline at $k = 1,200$ in the main experiment. All the approaches converge starting from this value, and not much effectiveness is achieved for higher values of k. In the generic experiment, both models achieve their first peaks at around $k = 700$. The performance of both models oscillates but keeps at the same rate for larger values of k. In the topic-specific experiments, the performance of $E_1^{0.005}$ increases significantly while approaching the value of $k = 400$. After reaching this value, the performance of $E_1^{0.005}$ converges, and no significant gain is achieved afterward. In comparison, the performance of E_2 converges after reaching the value of $k = 800$. The performance of both models, however, remains equal or less than that of the baseline.

Figure 5.5 (b, d, f) shows the effectiveness of our approach in the three experiments: main, topic-specific, and generic in the LSA space, respectively. As depicted in the three figures, the model $E_1^{0.005}$ outperforms E_2 in all cases, which shows that content-based topic-removal of arguments is more effective than using its structure. In the generic experiment, all models in the LSA space show subpar effectiveness compared to their counterpart in the TF-IDF space and lack clear peaks. In the topic-specific experiment, our approach outperforms the LSA baseline and their counterpart in the TF-IDF space. Nevertheless, like in the generic experiment, no clear peak is reached by any model.

Discussion

The results show the merit of removing topic-specific words of an argument for identifying its frame. According to the reported results, our approach is effective at identifying generic frames and does not suit identifying topic-specific frames. An interesting finding is that the premise of an argument carries more information about its frame than the conclusion. This is shown in the higher effectiveness achieved after applying E_2 compared to the baseline. A justification can be that a conclusion is more likely to carry stance-taking words toward the topic. In general, $E_1^{0.005}$ achieved higher results than E_2, which shows that using the content of an argument is more effective than using its structure to model frames. A possible justification for this can be that E_2 is more "aggressive" than needed at removing topic-specific features.

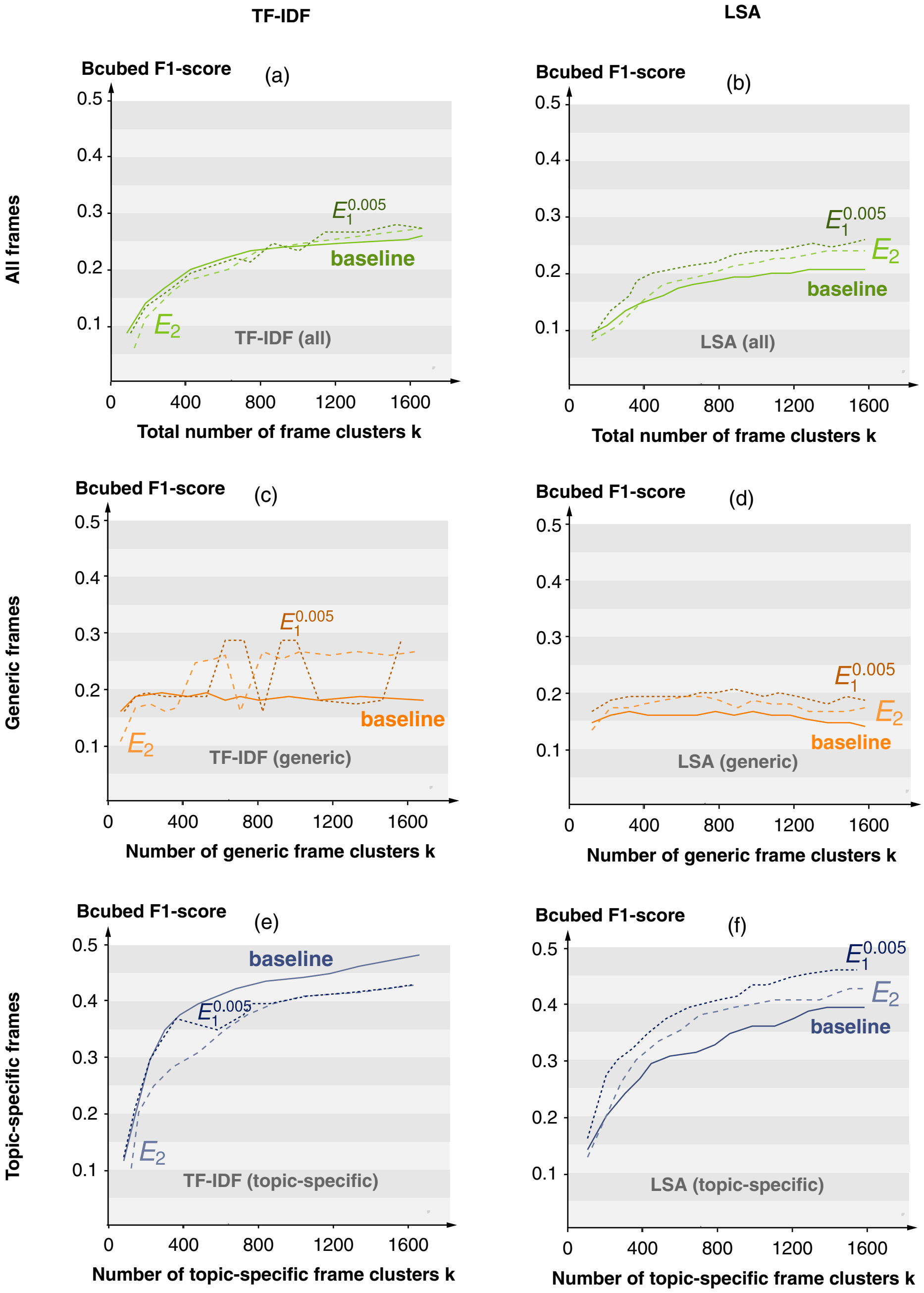


FIGURE 5.5: Effectiveness of frame clustering with TF-IDF and LSA without topic removal (baseline) and after applying $E_1^{0.005}$ and E_2 in (a, b) the main experiment, (c, d) the generic experiment, and (e, f) the topic-specific experiment.

5.1.5 Error Analysis

We analyze the topic and frame clusters produced by our approach to convey to the reader a sense of its performance. For topic clusters, we focus on the semantic space LSA Debate since our approach performed the best in this semantic space. For frame clusters, we analyze the output of our approach in the semantic space TF-IDF after applying $E_2^{0.005}$ since our approach performed the best in this semantic space. Our goal is to identify the topics and frames in the dataset that our approach completely confused or correctly identified. To identify these cases, we sort the topics and frames in the ground-truth dataset according to the maximum F1-score achieved in the aforementioned semantic spaces respectively. We manually analyze the topic and frames labels and the count of arguments they comprise and report the most interesting cases.

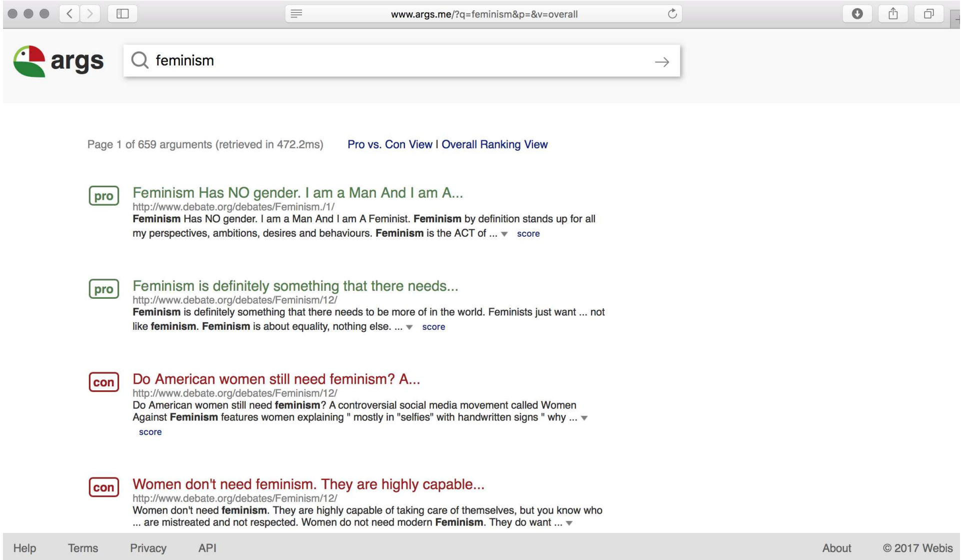
For topic clustering, examples of topics that our approach correctly identified (with an F1-score of 1) are: “zoos” and “compulsory vaccination”. On the other hand, our approach struggled at identifying topics like “Is Pluto a planet?” and “Immunity from prosecution for politicians’ (with an F1-score lower than 0.1). A reason for this might be that these topics are too specific and not covered well in our dataset.

In frame clustering, the hardest cases for our approach in the TF-IDF space were topic-specific frames that contain few arguments, e.g., “child disability”. Generic frames such as “rights” and “feasibility” were also hard to identify (with an F1-score lower than 0.1). A possible explanation is that these frames can be confused with generic frames like “human rights” and “economics”. Examples of generic frames that were effectively identified are “freedom of speech” and “public health” (with an F1-score equals to 0.5).

The analysis shows the challenges posed by the frame identification task and our dataset’s limitations. One of the main limitations of our approach and dataset is the assumption that frames are non-overlapping sets of arguments. A more realistic approach should consider cases where an argument emphasizes multiple frames instead of focusing on its primary frame. In the next section, we introduce a visual interface for delivering arguments in a retrieval scenario with the aspects they emphasize. The visual interface relaxes the aforementioned assumption by modeling an argument as a vector of numerical weights that correspond to multiple aspects.

5.2 Explorative Visualization of Argument Search Results

First prototypes of argument retrieval systems [151, 173] presented arguments in textual form with linked sources, similar to the web page snippets of conventional search engines, but with color-encoded stances. An example interface is given below in Figure 5.6 for args.me. This is adequate for comprehending those arguments deemed most relevant. Unlike many general information needs [42], how-



Page 1 of 659 arguments (retrieved in 472.2ms) Pro vs. Con View | Overall Ranking View

pro Feminism Has NO gender. I am a Man And I am A...
http://www.debate.org/debates/Feminism./1/
Feminism Has NO gender. I am a Man And I am A Feminist. **Feminism** by definition stands up for all my perspectives, ambitions, desires and behaviours. **Feminism** is the ACT of ... ▾ score

pro Feminism is definitely something that there needs...
http://www.debate.org/debates/Feminism/12/
Feminism is definitely something that there needs to be more of in the world. Feminists just want ... not like **feminism**. **Feminism** is about equality, nothing else. ... ▾ score

con Do American women still need feminism? A...
http://www.debate.org/debates/Feminism/12/
Do American women still need **feminism**? A controversial social media movement called Women Against **Feminism** features women explaining " mostly in "selfies" with handwritten signs " why ... ▾ score

con Women don't need feminism. They are highly capable...
http://www.debate.org/debates/Feminism/12/
Women don't need **feminism**. They are highly capable of taking care of themselves, but you know who ... are mistreated and not respected. Women do not need modern **Feminism**. They do want ... ▾

FIGURE 5.6: The *overall ranking view* of the initial version of args.me, showing results for the query "feminism".

ever, reading the top results is not enough for building an informed stance. Rather, diverse aspects of a controversial topic need to be explored. A recent study shows that comprehension and completeness are the most important factor of useful answers to non-factual questions after their relevance [34].

In this section, we introduce a novel way of presenting argument search results to support an aspect-guided exploration of the arguments on a topic. In particular, we visualize this *topic space* in a barycentric coordinate system [135], representing the distribution of pro and con arguments over the main covered aspects (see Figure 5.8). Possible aspects were derived offline from the Wikipedia list of controversial topics[3] as well as from Latent Dirichlet allocation (LDA) topic models built based on the 291,000 arguments in args.me index, whereas the aspects actually visualized are derived ad-hoc from the search results. Through interactions with the visualization, a user can easily highlight and filter arguments on the aspects of interest. In two case studies, we demonstrate how the visualization speeds up argument search notably.

5.2.1 Visualization of the Aspect Space

Determining Aspects

The first step to develop the visualization was to build a topic model that can represent the aspects of each argument in the result list. We compared two alternative

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

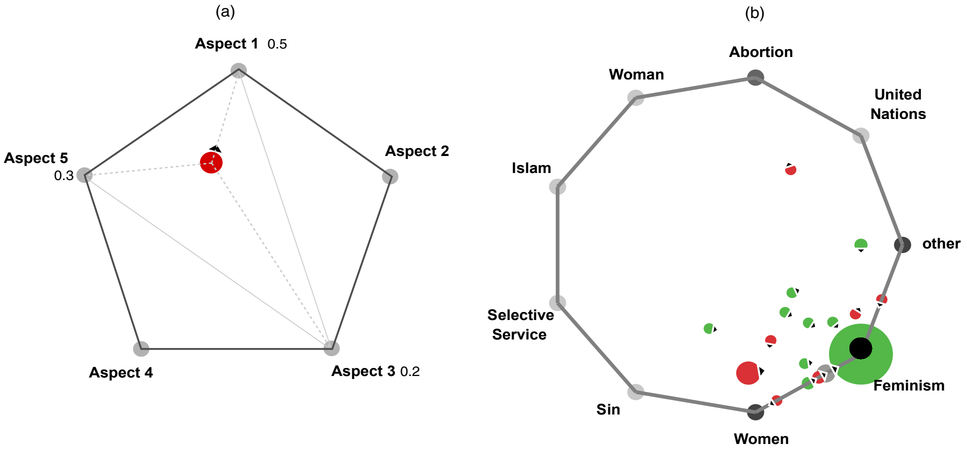


FIGURE 5.7: (a) Positioning an argument glyph in the topic space: the black arrow shows the linear combination of weighted vertices ("aspect 3" 0.2, "aspect 5" 0.3, "aspect 1" 0.5). The glyph itself points to the main covered aspect. (b) Topic space visualization for the query "feminism", positioning the retrieved arguments according to the eight main covered aspects and "other".

approaches for this purpose:

First, we computed the relative distribution of all the over 1,000 terms from the Wikipedia list of controversial topics in each indexed argument. For instance, if "women" occurs ten times, "woman" six times, "feminism" four times, and no other term, then we have ("women" 0.5, "woman" 0.3, "feminism" 0.2) with implicit zeros for all others. Second, we performed LDA topic modeling [25] based on the words in all arguments from our index. With an interval size of 10, we tested all numbers of topics from 10 to 100 and chose the number that minimized perplexity: 40. Each aspect is then represented by all words of one LDA topic, and the relative aspect distribution is calculated by counting the occurrence of all associated words in each argument. We found the Wikipedia-based topic model to be more convincing, which is why it is set as the default in args.me.[4]

Visualizing Aspects

To visualize the aspect-based topic space, we opted for generalized barycentric coordinates [100], as they naturally fit our purpose: We represent an argument as a linear combination of weights for all aspects, while barycentric coordinates represent a point as a linear combination of the vertices of a polygon (both adding up to 1.0).[5] Thus, the topic model can be used as input for the visualization without recalculation. Figure 5.7 (b) shows the visualization of the results for the query

[4] The LDA alternative can be activated in args.me by changing the value of the v-parameter in the URL field to "lda".

[5] Dora Kiesel, Patrick Riehmann, and Bernd Fröhlich proposed the idea of using generalized barycentric coordinates to visualize the topic space as well as its design.

“feminism”, consisting of two main elements: the *topic space* and the *argument glyphs* within this space.

The topic space is depicted as a regular polygon with one vertex for each represented aspect. Both given topic model alternatives comprise too many aspects to depict them all. To reduce visual clutter in favor of a lean visualization, we limit the maximum number of visualized aspects so that readability is not diminished. In particular, we keep only those eight aspects that are the most frequent in the argument search results. All other aspects are merged into a ninth aspect “other”. The labels for the aspects are short terms in case of the Wikipedia-based topic model or visualized as word clouds in case of the LDA topic model.

Each argument glyph represents one argument in the form of a colored circle (green for pro, red for con) with a small arrow pointing to the main covered aspect. The glyphs are positioned based on their aspect distribution: the stronger one aspect, the stronger a glyph is “pulled” in that direction, as sketched in Figure 5.7. Accordingly, similar arguments are placed spatially near each other. To ensure the visibility of all glyphs and to avoid overplotting, arguments placed on top of each other are aggregated into a single glyph. The glyph size depends on a logarithmic mapping of the number of represented arguments. Since arguments with both stances may be grouped, the color of an aggregate glyph represents the majority stance of all arguments contained, from green (all pro), over gray (balanced pro/con), to red (all con).

Interacting with Aspects

The integration of our visualization into args.me is shown below in Figures 5.9 and 5.10. This new *topic space view* replaces the old overall ranking view: it includes the textual argument ranking and adds the visualization to the right. At first, the visualization shows only the information outlined above, but it provides further details upon interaction.

Barycentric coordinates are ambiguous and may place arguments with different aspects at similar locations. For disambiguation, users can hover over a glyph to reveal all covered aspects, as exemplified in Figure 5.8 (a). The represented arguments are also highlighted in the textual list, given that they appear on the current result page. Vice versa, hovering over a textual argument highlights the respective glyph with a wide green or red border.

In addition, the visualization allows users to filter the textual results: A user can select one or more arguments by clicking or brushing (see Figure 5.8 (b)), in order to narrow down the list to the aspects of interest. All other arguments are grayed out.

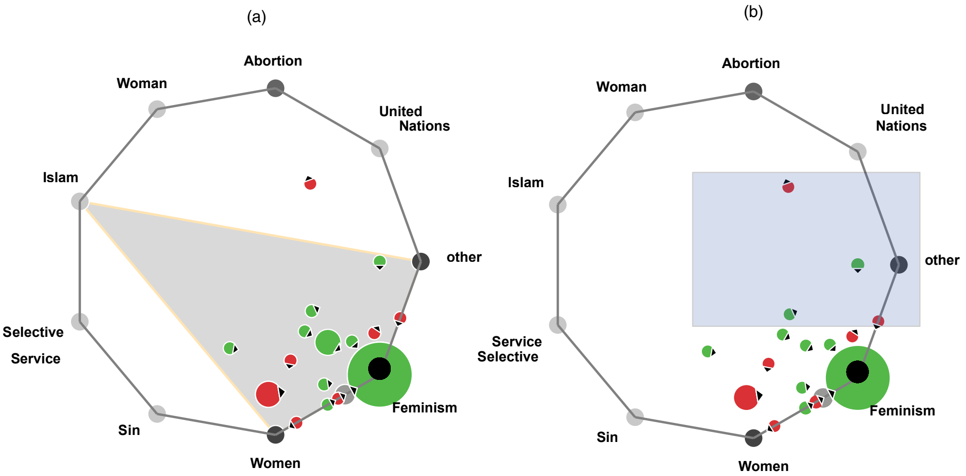
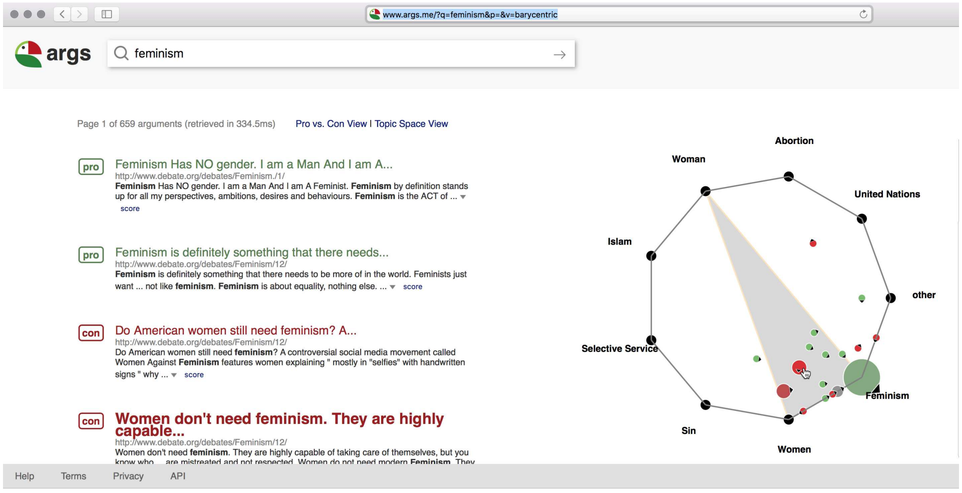


FIGURE 5.8: Interacting with the visualization through: a) Hovering over an argument reveals the aspects it covers (main aspect marked by a small arrow). b) Selecting arguments in the topic space visualization filters them in the textual result list of args.me.



Page 1 of 659 arguments (retrieved in 334.5ms) Pro vs. Con View I Topic Space View

pro Feminism Has NO gender. I am a Man And I am A...
http://www.debate.org/debates/Feminism./1/
Feminism Has NO gender. I am a Man And I am A Feminist. **Feminism** by definition stands up for all my perspectives, ambitions, desires and behaviours. **Feminism** is the ACT of ... ▾
score

pro Feminism is definitely something that there needs...
http://www.debate.org/debates/Feminism/12/
Feminism is definitely something that there needs to be more of in the world. Feminists just want ... not like **feminism**. **Feminism** is about equality, nothing else. ... ▾ score

con Do American women still need feminism? A...
http://www.debate.org/debates/Feminism/12/
Do American women still need **feminism**? A controversial social media movement called Women Against **Feminism** features women explaining " mostly in "selfies" with handwritten signs " why ... ▾ score

con **Women don't need feminism. They are highly capable...**
http://www.debate.org/debates/Feminism/12/
Women don't need **feminism**. They are highly capable of taking care of themselves, but you know who ... are mistreated and not respected. Women do not need modern **Feminism**. They

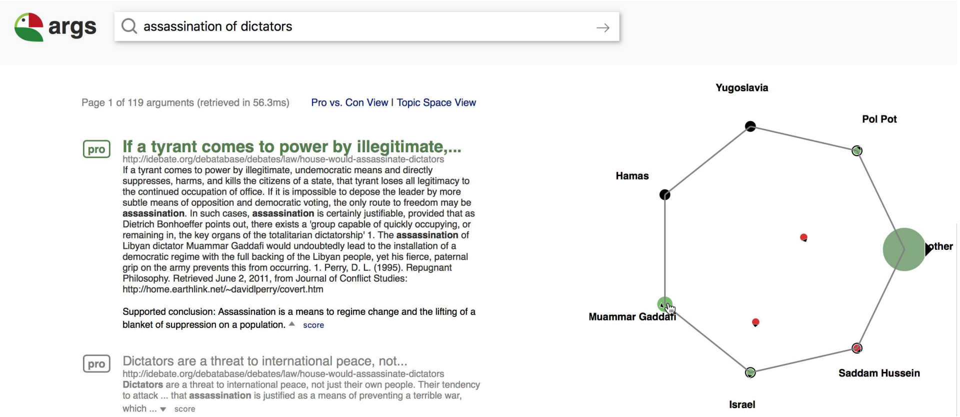
FIGURE 5.9: The args.me search results for the query “feminism”, along with the integrated topic space visualization. The argument hovered over in the visualization is highlighted in the textual result list (“Women don’t need...”).

5.2.2 Case Studies

To verify the benefit of our visualization, we finally explore two typical use cases of argument search: *topic space exploration* and *search refinement*.

Topic Space Exploration

First, we consider a query for “feminism”. Args.me returns 659 arguments for this topic, as shown in Figure 5.9. While the top-ranked arguments seem highly relevant in general, our visualization reveals that also some rather specific aspects are



assassination of dictators

Page 1 of 119 arguments (retrieved in 56.3ms) Pro vs. Con View I Topic Space View

pro **If a tyrant comes to power by illegitimate,...**
http://idebate.org/debatabase/debates/law/house-would-assassinate-dictators
If a tyrant comes to power by illegitimate, undemocratic means and directly suppresses, harms, and kills the citizens of a state, that tyrant loses all legitimacy to the continued occupation of office. If it is impossible to depose the leader by more subtle means of opposition and democratic voting, the only route to freedom may be **assassination**. In such cases, **assassination** is certainly justifiable, provided that as Dietrich Bonhoeffer points out, there exists a 'group capable of quickly occupying, or remaining in, the key organs of the totalitarian dictatorship' 1. The **assassination** of Libyan dictator Muammar Gaddafi would undoubtedly lead to the installation of a democratic regime with the full backing of the Libyan people, yet his fierce, paternal grip on the army prevents this from occurring. 1. Perry, D. L. (1995). Repugnant Philosophy. Retrieved June 2, 2011, from Journal of Conflict Studies: http://home.earthlink.net/~davidlperry/covert.htm

Supported conclusion: Assassination is a means to regime change and the lifting of a blanket of suppression on a population. ▲ score

pro Dictators are a threat to international peace, not...
http://idebate.org/debatabase/debates/law/house-would-assassinate-dictators
Dictators are a threat to international peace, not just their own people. Their tendency to attack ... that **assassination** is justified as a means of preventing a terrible war, which ... ▼ score

FIGURE 5.10: The single filtered args.me search result on the aspect "Muammar Gaddafi" for the query "assassination of dictators". The filtering is the result of clicking on the respective argument glyph in the topic space visualization.

covered by the search results, such as "abortion" and "United Nations". Interacting with the visualization helps to explore the entire topic space.

In particular, hovering over the argument glyphs clarifies what aspects they exactly cover, such as "woman", "women", and "feminism" for the highlighted argument in Figure 5.9. After a first exploration via hovering, a result subset of interest can be filtered through brushing, say, the four top-most glyphs (see Figure 5.8 (b) above). The selected arguments are then shown at the top of the textual result list (all below are grayed out). From the selected arguments, we learn that Emma Watson has made the need for feminism a point at the United Nations, whereas the claimed necessity of abortion is used as an analogy to justify the necessity of feminism. Without the visualization, these insights would have been hard to gain; the two respective arguments were ranked at positions #43 and #46.

Search Refinement

As a second example, we assume that a user looks for new arguments on the "assassination of dictators", for which args.me provides 119 results. If the user wants to refine a search to restrict it to a specific aspect of the topic only (e.g., to arguments covering "Muammar Gaddafi"), a simple click on the respective argument glyph in the topic space visualization suffices, as illustrated in Figure 5.10. The associated arguments are filtered and placed at the top of the result list (only one argument in the illustrated case). With the existing interaction methods of args.me, the argument text can be extended, and its source web page shows up after clicking on it. In the old overall ranking view, the argument would have been ranked at position #34.

5.3 Summary

In this chapter, we made several contributions along the way of enabling users to find arguments based on the frames they capture. First, we introduced an approach to identify the frames of an argument and then proposed a visual interface to present and explore the retrieved arguments by their aspects.

To identify the frames of an argument, we proposed an unsupervised approach that takes a set of arguments as input and clusters them into frames. The approach consists of three steps: it groups similar arguments into clusters, removes topic-specific tokens in each cluster, and then clusters the arguments once again into frames. To evaluate the approach, we introduced a dataset of 12,326 arguments on 465 topics which are labeled with 1,623 frames. The frames in the dataset cover into 330 generic frames and 1,293 topic-specific frames. Whereas generic frames are used for more than one topic, topic-specific frames are used only for one topic. We conducted experiments to evaluate our approach at identifying generic frames and topic-specific frames independently and together. The experiments show that removing topic-specific features from arguments benefits identifying generic frames.

Delivering arguments with their frames in a search engine needs a dedicated interface that presents the retrieved arguments along the frames. In Section 5.2, we proposed a visual interface to explore the retrieved arguments for a query by the aspects they cover. We modeled an argument in the retrieved arguments list as a set of weighted aspects where the sum of the weights is 1. The retrieved arguments to a query are then mapped to space whose coordinates are the top eight aspects of the retrieved arguments. In two use cases, we showed how the visual interface enables users to explore or refine the arguments in an efficient way by using functionalities provided by the visual interface (e.g., filtering arguments by an aspect).

Chapter 6

Conclusion

Web search engines are effective at answering questions that look for facts. However, they struggle at delivering all perspectives on information needs related to forming an opinion on a controversial topic. In this thesis, we aimed at developing methods to enable frame-guided retrieval of arguments in the context of search engines. To this end, we started by developing methods to identify information needs related to forming opinions on a controversial topic in the query stream of a search engine. Exploiting the web to answer these information needs requires argument mining approaches that can generalize across genre. In Chapter 4, we aimed at developing an approach that can detect argument units in natural text in a genre-robust way. Apart from genre, argument mining approaches should generalize to topics that are issued in a search context, which might not be covered by the corpora on which they are trained. Hence, we developed methods to foster the generalizability over topic while creating argument corpora and developing argument mining tasks. Arguments on a controversial topic cover different frames which are effective principles for selecting arguments that suit a target audience. In Chapter 5, we developed an approach to identify the frames of an argument and an interface to present the retrieved arguments together with the aspects they cover. In the following, we reintroduce the main research questions approached in this thesis and the main findings. Afterward, we present the limitations of the pursued approaches and possible future research directions.

6.1 Findings and Implications

Argument retrieval systems help users to form opinions on a controversial topic in an unbiased way. Analyzing and identifying information needs that seek to form an opinion is the first step of integrating argument retrieval systems into search engines. Existing research developed approaches to automatically distinguish opinion questions from factual and social questions in community question answering [37, 101]. In Chapter 3, we approached the research question RQ1: *How to identify argumentative questions in the context of search engines?* In comparison to questions posted in community question answering, questions asked on web search engines are more concise. This makes the task of identifying argumentative questions harder, since less information is provided about the question intent in web search engines.

To answer RQ1, we sampled a dataset from Yandex query logs that covers 19 controversial topics. In a crowdsourcing study, we annotated each question in the dataset with whether it is about a controversial topic or not, and if so, whether it seeks a fact, methods, or arguments. The crowdsourcing study shows that 28% of questions on controversial topics are argumentative. This clearly speaks for the importance of developing dedicated methods to handle this type of questions. A comparative analysis of the argumentative question types against the other question types shows that argumentative questions mainly ask for reasons and predictions. We realized an automatic classification of the three question types in the dataset using a supervised classifier. The classifier achieved high effectiveness (macro F1-score of 0.78) on the task of classifying questions in the dataset, even on unseen topics. The high classification performance opens the way for measuring and controlling how search engines influence public opinion on controversial topics. Also, it triggers a realistic adaptation of argument retrieval approaches in web search engines that goes beyond answering imagined queries to real questions that users ask.

The web offers the broadest argumentative content to answer argumentative questions. The automatic extraction of arguments on the web is challenged by the different genres the web covers (e.g., news or debate portals). Most existing argument mining approaches are developed for single genres and assume argument units to span a specific syntactic unit (e.g., sentence or clause). Automatic segmentation of text into argument units was approached only in argumentative essays by Stab [154]. Mining arguments on the web requires a generic segmentation method that is effective across genre.

In Chapter 4, we raised the research question RQ2: *How to extract argument units in a genre-robust way?* To this end, we framed the argument unit segmentation task on the token level and conducted two types of experiments: in-genre and cross-genre. In these experiments, we carried out the first systematic analysis of linguistic features and sequence-to-sequence models that take an increasing length of context while deciding the label of a token. In contrast to the approach of Stab [154], our approaches are evaluated on and across multiple genres and achieve a higher F1-score on the Essays corpus (0.89 in comparison to 0.87 of Stab [154]). The reason for the increased performance is the usage of Bi-LSTM, which incorporates a broad context while detecting argument units in a document. The results of the experiments illustrate that structural and semantic features are the most effective for argument unit segmentation across genres, while semantic features are the best for detecting the boundaries of argumentative units within genres. The results also demonstrate that while our approach for argument unit segmentation is effective in an in-genre setting, more knowledge is needed to achieve a cross-genre transfer. This shows that the established order of argument mining steps (unit segmentation and then unit type classification) for an in-genre setting can not be transferred into a cross-genre setting.

Argument mining approaches should be able to generalize to topics beyond those on which they are trained. A research question that we investigated in this direction is RQ3: *How to assess and foster generalizability over topic in argument mining approaches?* Our first contribution to answering this research question is a literature review of 45 argument corpora with respect to how many controversial topics they cover and how the topics were chosen. The literature review shows that the majority (about 65%) of argument corpora either cover up a small set of topics (up to 25 topics) or are provided with no topic labels. To further assess the overlap and coverage of the corpora's topics, we identified three authoritative topic ontologies that we introduced to the research community. Assessing the topic coverage of argument corpora was done manually by mapping the topics of argument corpora to the topic ontology. To extend this analysis to argument corpora that miss explicit topic labels, we developed approaches to automatically identify the topic of an argumentative document. This analysis showed that argument corpora which are provided with topic labels cover only a subset of the ontology topics. We also found that the distribution of the covered topics is heavily skewed toward a small subset. This clearly shows that existing argument corpora are biased toward a small set of topics.

Topic bias exerted in argument corpora renders approaches developed on them ungeneralizable to new topics. The topic-dependence of computational argumentation approaches should be taken into account, starting from acquiring corpora to developing experiments and applications. Creation of future argument corpora should be created with clear topic selection criteria with regard to an accepted topic source (e.g., one of the presented topic ontologies). An important implication of our work also is that the relatedness between topics should be taken into account while designing experiments that evaluate the generalizability of argument mining approaches.

Apart from corpus creation standards, fostering generalizability to new topics should be tackled in how argument mining tasks are formulated. Stance classification is the task of classifying an argument into pro or con for a given topic. To tackle the topic dependence of this task, we introduced the same side stance classification task, which takes a pair of arguments as input and returns whether they are on the same or the opposite side. The task was run as a shared task, which is based on an argument dataset that we sampled from the args.me corpus on two topics, and which we released to the research community. To assess the topic transfer, we ran within-topic experiments and cross-topic experiments and evaluated the approaches of nine participants.

The results of the same side stance classification experiments demonstrate that the performance of transformer-based approaches is similar in the within-topic and cross-topic experiments. Transformer-based approaches suffer a drop of only five to six accuracy points between the within-topic and cross-topic experiments, showing the feasibility of a topic-agnostic approach for stance classification. This also illustrates that detecting the stances of a pair of arguments in a comparative way is realizable and fosters cross-topic robustness. Apart from cross-topic robustness, the results of the experiments show the importance of processing multiple arguments on the same topic at the same time instead of single arguments. Processing multiple arguments simplify the stance classification task by providing a classifier with more knowledge about the topic. An advantageous property that characterizes same side stance classification is its adaptability to different and possibly even fine-granular stance label schemes (e.g., pro, con, and neutral).

To be effective in argumentation, an author should pick frames that resonate with the target audience. Delivering arguments with the frames they capture requires methods to identify the frames of an argument. For this goal, we raised the research question RQ4: *How to model and identify frames of an argument?* A key contribution in this regard is conceptualizing a frame as a set of arguments that shares an aspect and framing as selecting a set of arguments on a topic. We operationalized the model in a three-step approach: The first step groups arguments according to their topic similarity. The second step removes topic-specific tokens by exploiting the topic clusters or the argument structure. Finally, our approach clusters the topic-free arguments into frames. To evaluate the approach, we introduced to the research community the first framing dataset that is labeled with topic-specific and generic frames. By targeting topic-specific, generic, and all frames in different experiments, we showed that removing topic-specific features helps to identify frames. Particularly, we found that identifying generic frames benefits from removing topic features, which are actually the hardest case. On the other hand, removing topic features cannot help in identifying topic-specific frames.

Our study sets a lacking methodology for modeling frames in argumentation that is based on a formal definition of frames. The methodology can be extended with different approaches for topic identification and removal as well as argument clustering. The main implication of our study is showing the importance of discerning the topic of an argument in order to identify how the argument frames the topic. We expect future approaches to utilize structured knowledge (e.g., topic ontologies) to identify the topic and frames of arguments.

Existing interfaces for argument retrieval present arguments as a list of snippets sorted into pro and con. A dedicated interface is needed to present the frames for a controversial topic along with the arguments that capture each of the frames.

At the end of Chapter 5, we introduced an interface that presents arguments as groups mapped into an aspect space. In two use cases, we showed how the interface enables an efficient way of exploring the arguments by the aspects they cover.

6.2 Future Work

Chapter 3 introduces the first study of argumentative questions in the context of web search. That being said, a key limitation of the analysis lies in the manual selection of the controversial topics and their descriptors. The 19 controversial topics that were selected for the analysis included global issues that are mostly debated on online debate portals or local issues that were trending in Russia in 2012. While selecting these topics enabled us to develop the first approach to identify argumentative questions and their characteristics, an open challenge remains to scale up our approach to cover a wider range of controversial topics. Applying our approach to a broader set of topics requires developing approaches that classify whether a question is about a controversial topic, which is a research area on its own in question answering research.

Future research on argumentative questions should distinguish those questions that have direct consequences on the lifes of the askers, their close social circle, and their environments. Whereas we focused in this thesis on controversial topics that are relevant to society (e.g., “Nord Stream 2”), users are likely to ask more argumentative questions that are relevant to their private life. Search engines should carefully treat argumentative questions that affect the health of the asker (e.g., deciding on a medical operation), their social status (e.g., “studying” or “working abroad”), and their relations with family members (e.g., adoption or divorce). Since these questions might have drastic consequences on the lifes of the askers, handling these argumentative questions in an unbiased way is an ethical obligation of search engines.

The results of our experiments on the generalizability of argument unit segmentation over genre warrants further inquiry. A key observation that future research should take into account is that the length, structure, and position of premises are different from conclusions. This speaks for performing argument unit segmentation jointly with unit type classification (into premise or conclusion) or after it. Another research direction is to assess the generalizability of argument unit segmentation over genre in a topic-specific way.

In Section 4.2, we introduced three authoritative topic ontologies and used them to assess topic bias in argument corpora. While the study elucidates the problem of topic bias in argument corpora, we noticed an overlap between the topics covered by the three topic ontologies. Fostering the usage of topic selection guidelines in constructing argument corpora requires having one accepted topic ontology that compromises all recognized controversial topics. Hence, future research can work on unifying the three topic ontologies by merging semantically similar topics together. Our approach for identifying the topic of arguments achieved moderate effectiveness because of the large space of controversial topics (about 742 for Wikipedia). Future research can improve upon our approach by utilizing the structure of the topic ontology using hierarchical classifiers. Hierarchical classifiers first map a document to one topic in the upper level and then consider only the subtopics of this topic for classification in the lower levels. In this way, the space of controversial topics in the lower levels can be largely reduced.

While same side stance classification fosters topic-robustness, we see several possibilities for future research. Extending the dataset to cover more topics is a logical next step. Another limitation of our approach is the assumption that same side stance classification presupposes, which is that the input arguments are on the same topic. Whereas such an assumption can be made in a retrieval scenario, topic identification or filtering approaches need to be used to apply same side stance classification in different applications. Another possible research direction is extending same side stance classification to return the stance similarity between a pair of arguments as a real value. This will allow revealing more fine-granular stance information of arguments. For example, grouping arguments with perfect stance similarity or identifying arguments with salient stances.

Chapter 5 introduced a lacking methodology for identifying frames in argumentation. Still, effective identification of frames requires developing more effective approaches for topic identification, modeling multiple frames in an argumentative text, and integrating relevant characteristics of the audience of an argument. Whereas we used the content of an argument to identify its topic, a clear improvement can be achieved by utilizing metadata provided with an argument (e.g., topic and stance labels) or external knowledge such as Wikipedia or topic ontologies.

One limitation of our dataset and approach for frame identification is the assumption that an argument covers only one primary frame. While this assumption allowed for a simple evaluation of our approach, a more general approach should consider cases where arguments emphasize multiple frames. Future research direction can develop approaches that assign multiple frames to an argument. A further interesting research direction is to develop models to capture frame patterns in an argumentative text. For example, a sequential model of frames can be used to reveal in which sequence frames are delivered and what influences the choice of the frames (e.g., audience).

The choice of frames depends largely on the target audience of the argumentative text. A possible research direction is to study how to model the audience in framing, i.e., what kind of information about the audience is relevant to the choice of frames. Such information can include known frames or topics, previous beliefs, as well as certain attributes of the audience (e.g., age).

Bibliography

[1] Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn Walker. 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), May 2016. URL `https://aclanthology.org/L16-1704`.

[2] Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 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, June 2014.

[3] Yamen Ajjour, Wei-Fan Chen, Johannes Kiesel, Henning Wachsmuth, and Benno Stein. Unit Segmentation of Argumentative Texts. In *Proceedings of the Fourth Workshop on Argument Mining*. Association for Computational Linguistics, 2017.

[4] Yamen Ajjour, Henning Wachsmuth, Dora Kiesel, Patrick Riehmann, Fan Fan, Giuliano Castiglia, Rosemary Adejoh, Bernd Fröhlich, and Benno Stein. Visualization of the topic space of argument search results in args.me. In Eduardo Blanco and Wei Lu, editors, *2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018) - System Demonstrations*, pages 60–65. Association for Computational Linguistics, November 2018. URL `http://aclweb.org/anthology/D18-2011`.

[5] Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. Modeling Frames in Argumentation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP 2019)*, pages 2922–2932. ACL, November 2019. URL `https://www.aclweb.org/anthology/D19-1290`.

[6] Yamen Ajjour, Henning Wachsmuth, Johannes Kiesel, Martin Potthast,

Matthias Hagen, and Benno Stein. Data Acquisition for Argument Search: The args.me corpus. In Christoph Benzmüller and Heiner Stuckenschmidt, editors, *42nd German Conference on Artificial Intelligence (KI 2019)*, pages 48–59, Berlin Heidelberg New York, September 2019. Springer. doi:10.1007/978-3-030-30179-8_4.

[7] Yamen Ajjour, Pavel Braslavski, Alexander Bondarenko, and Benno Stein. Identifying Argumentative Questions in Web Search Logs. In *45th International ACM Conference on Research and Development in Information Retrieval (SIGIR 2022)*. ACM, July 2022. doi:10.1145/3477495.3531864.

[8] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649, 2018.

[9] Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein. 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*, pages 1395–1404. Association for Computational Linguistics, 2016. URL `http://aclweb.org/anthology/N16-116`.

[10] Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. A News Editorial Corpus for Mining Argumentation Strategies. In Yuji Matsumoto and Rashmi Prasad, editors, *26th International Conference on Computational Linguistics (COLING 2016)*, pages 3433–3443. Association for Computational Linguistics, December 2016. URL `http://aclweb.org/anthology/C16-1324`.

[11] Khalid Al-Khatib, Viorel Morari, and Benno Stein. Style Analysis of Argumentative Texts by Mining Rhetorical Devices. In *the 7th Workshop on Argument Mining*, pages 106–116. ACL, December 2020. URL `https://aclanthology.org/2020.argmining-1.12`.

[12] Giambattista Amati. Frequentist and bayesian approach to information retrieval. In *European Conference on Information Retrieval*, pages 13–24, Berlin, Heidelberg, 2006. Springer.

[13] Enrique Amigó, Julio Gonzalo, Javier Artiles, and Fleisa Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. 12:461–486, 2009.

[14] Aristotle and George A. Kennedy. *On Rhetoric: A Theory of Civic Discourse*. Oxford: Oxford University Press, 2006.

[15] M. Azar. Argumentative Text as Rhetorical Structure: An Application of Rhetorical Structure Theory. *Argumentation*, 13:97–114, 1999.

[16] Leif Azzopardi. Cognitive Biases in Search: A Review and Reflection of Cognitive Biases in Information Retrieval. In Falk Scholer, Paul Thomas, David Elsweiler, Hideo Joho, Noriko Kando, and Catherine Smith, editors, *Proceedings of the Conference on Human Information Interaction and Retrieval (CHIIR 2021)*, pages 27–37. ACM, 2021. doi:10.1145/3406522.3446023. URL `https://doi.org/10.1145/3406522.3446023`.

[17] Amit Bagga and Breck Baldwin. Algorithms for scoring coreference chains. In *The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pages 563–566, 1998.

[18] Alexandra Balahur, Ester Boldrini, Andrés Montoyo, and Patricio Martínez-Barco. A Comparative Study of Open Domain and Opinion Question Answering Systems for Factual and Opinionated Queries. In *Proceedings of the International Conference RANLP-2009*, pages 18–22. Association for Computational Linguistics, 2009. URL `http://www.aclweb.org/anthology/R09-1004`.

[19] Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. Stance Classification of Context-Dependent Claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 251–261. Association for Computational Linguistics, 2017. URL `http://aclweb.org/anthology/E17-1024`.

[20] Roy Bar-Haim, Dalia krieger, Orith Toledo-Ronen, Lilach Edelstein, Yonatan Bilu, Alon Halfon, Yoav Katz, Amir Menczel, Ranit Aharonov, and Noam Slonim. 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, 2019. URL `https://www.aclweb.org/anthology/P19-1094/`.

[21] Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. From Arguments to Key Points: Towards Automatic Argument Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4029–4039. Association for Computational Linguistics, July 2020. doi:10.18653/v1/2020.acl-main.371. URL `https://www.aclweb.org/anthology/2020.acl-main.371`.

[22] Floris Bex, John Lawrence, Mark Snaith, and Chris Reed. Implementing the Argument Web. *Communications of the ACM*, 56:66–73, 10 2013. Crawled in Jan, 2020.

[23] Yonatan Bilu, Ariel Gera, Danel Hershcovich, Benjamin Sznajder, Dan Lahav, Guy Moshkowich, Anael Malet, Assaf Gavron, and Noam Slonim. 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, 2019. URL `https://www.aclweb.org/anthology/P19-1097/`.

[24] J. Anthony Blair. *Groundwork in the Theory of Argumentation*. Springer Netherlands, 2012.

[25] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003. URL `http://dl.acm.org/citation.cfm?id=944919.944937`.

[26] Filip Boltužić and Jan Šnajder. 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, 01 2014.

[27] Alexander Bondarenko, Pavel Braslavski, Michael Völske, Rami Aly, Maik Fröbe, Alexander Panchenko, Chris Biemann, Benno Stein, and Matthias Hagen. Comparative Web Search Questions. In James Caverlee, Xia (Ben) Hu, Mounia Lalmas, and Wei Wang, editors, *13th ACM International Conference on Web Search and Data Mining (WSDM 2020)*, pages 52–60. ACM, February 2020. URL `https://dl.acm.org/doi/abs/10.1145/3336191.3371848`.

[28] Alexander Bondarenko, Maik Fröbe, Meriem Beloucif, Lukas Gienapp, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. Overview of Touché 2020: Argument Retrieval. In Linda Cappellato, Carsten Eickhoff, Nicola Ferro, and Aurélie Névéol, editors, *Working Notes Papers of the CLEF 2020 Evaluation Labs*, volume 2696 of *CEUR Workshop Proceedings*, September 2020. URL `http://ceur-ws.org/Vol-2696/`.

[29] Alexander Bondarenko, Lukas Gienapp, Maik Fröbe, Meriem Beloucif, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. Overview of Touché 2021: Argument Retrieval. In Djoerd Hiemstra, Maria-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani,

editors, *Advances in Information Retrieval. 43rd European Conference on IR Research (ECIR 2021)*, volume 12036 of *Lecture Notes in Computer Science*, Berlin Heidelberg New York, March 2021. Springer.

[30] Alexander Bondarenko, Yamen Ajjour, Valentin Dittmar, Niklas Homann, Pavel Braslavski, and Matthias Hagen. Towards Understanding and Answering Comparative Questions. In K. Selcuk Candan, Huan Liu, Leman Akoglu, Xin Luna Dong, and Jiliang Tang, editors, *Proceedings of the 15th ACM International Conference on Web Search and Data Mining (WSDM 2022)*, pages 66–74. ACM, February 2022. doi:10.1145/3488560.3498534. URL `https://dl.acm.org/doi/10.1145/3488560.3498534`.

[31] Amber E. Boydstun, Dallas Card, Justin H. Gross, Philip Resnik, and Noah A. Smith. Identifying media frames and frame dynamics within and across policy issues. In *Proceedings of the Workshop on New Directions in Analyzing Text as Data*, 2013.

[32] Elena Cabrio and Serena Villata. Natural Language Arguments: A Combined Approach. In *Proceedings of the 20th European Conference on Artificial Intelligence*, pages 205–210. IOS Press, 2012.

[33] Elena Cabrio and Serena Villata. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 208–212, 2012.

[34] B. Barla Cambazoglu, Valeriia Baranova, Falk Scholer, Mark Sanderson, Leila Tavakoli, and Bruce Croft. Quantifying Human-Perceived Answer Utility in Non-Factoid Question Answering. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval (CHIIR 2021)*, pages 75–84. ACM, 2021. URL `https://doi.org/10.1145/3406522.3446028`.

[35] Dallas Card, Amber E Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. The Media Frames Corpus: Annotations of Frames Across Issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 438–444. Association for Computational Linguistics, 2015.

[36] Sergiu Chelaru, Ismail Sengör Altingövde, Stefan Siersdorfer, and Wolfgang Nejdl. Analyzing, Detecting, and Exploiting Sentiment in Web Queries. *ACM Trans. Web*, 8(1):6:1–6:28, 2013. doi:10.1145/2535525. URL `https://doi.org/10.1145/2535525`.

[37] Long Chen, Dell Zhang, and Mark Levene. Understanding User Intent in Community Question Answering. In Alain Mille, Fabien Gandon, Jacques Misselis, Michael Rabinovich, and Steffen Staab, editors, *Proceedings of the 21st World Wide Web Conference (WWW 2012)*, pages 823–828. ACM, 2012. doi:10.1145/2187980.2188206. URL https://doi.org/10.1145/2187980.2188206.

[38] Sachin Pathiyan Cherumanal, Damiano Spina, Falk Scholer, and W. Bruce Croft. Evaluating fairness in argument retrieval. In Gianluca Demartini, Guido Zuccon, J. Shane Culpepper, Zi Huang, and Hanghang Tong, editors, *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, pages 3363–3367. ACM, 2021. doi:10.1145/3459637.3482099. URL https://doi.org/10.1145/3459637.3482099.

[39] Michael Collins. Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, volume 10, pages 1–8. Association for Computational Linguistics, 2002.

[40] Alexander Conard, Janyce Wiebe, and Rebecca Hwa. 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, July 2012.

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

[42] Bruce Croft, Donald Metzler, and Trevor Strohman. *Search Engines: Information Retrieval in Practice*. Addison-Wesley, USA, 1st edition, 2009.

[43] Edward Damer. Attacking faulty reasoning: A practical guide to. 2009.

[44] Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 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, 2017. URL https://www.aclweb.org/anthology/D17-1218/.

[45] Scott Deerwester, Susan T. Dumals, George W. Furnasand, Thomas K Landauer, and Richard Harshman. Indexing by Latent Semantic Analysis. 41 (6):391–407, 1990.

[46] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL `http://arxiv.org/abs/1810.04805`.

[47] Lorik Dumani and Ralf Schenkel. Quality Aware Ranking of Arguments. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM 2020)*, pages 335–344. Association for Computing Machinery, 2020. URL `https://doi.org/10.1007/978-3-030-45439-5_29`.

[48] Lorik Dumani, Patrick J. Neumann, and Ralf Schenkel. A framework for argument retrieval - ranking argument clusters by frequency and specificity. In *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part I*, volume 12035, pages 431–445. Springer, 2020. URL `https://doi.org/10.1007/978-3-030-45439-5_29`.

[49] Frans Van Eemeren and Peter Houtlosser. Strategic Manoeuvring in Argumentative Discourse. *Discourse Studies*, 1(4):479–497, 1999. doi:10.1177/1461445699001004005.

[50] Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. Neural End-to-End Learning for Computational Argumentation Mining. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017. In Press.

[51] 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. 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, 2020. URL `https://aaai.org/ojs/index.php/AAAI/article/view/6270`.

[52] Robert M. Entman. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58, 1993.

[53] Shnarch Eyal, Leshem Choshen, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. 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, November 2020. doi:10.18653/v1/2020.findings-emnlp.243. URL https://www.aclweb.org/anthology/2020.findings-emnlp.243.

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

[55] William Ferreira and Andreas Vlachos. Emergent: A Novel Data-set for Stance Classification. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 1163–1168. The Association for Computational Linguistics, 2016. doi:10.18653/v1/n16-1138. URL https://doi.org/10.18653/v1/n16-1138.

[56] Anjalie Field, Doron Kliger, Shuly Wintner, Jinnifer Pan, Dan Jurafsky, and Yulia Tesvetkov. Framing and Agenda-setting in Russian News: a Computational Analysis of Intricate Political Strategies, 2018.

[57] Michael Fromm, Evgeniy Faermann, and Thomas Seidl. TACAM: Topic And Context Aware Argument Mining. In *The proceedings of the 2019 International Conference on Web Intelligence (IEEE/WIC/ACM 2019)*, pages 99–106, 2019.

[58] Jonas Gabrielsen, Sine NØrholm Just, and Mette Bengtsson. Concepts and Contexts – Argumentative Forms of Framing. *Proceedings of the 7th Conference of the Society for the Study of Argumentation*, pages 533–543, 2011.

[59] Evgeniy Gabrilovich and Shaul Markovitch. 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, 2007. URL http://ijcai.org/Proceedings/07/Papers/259.pdf.

[60] Gizem Gezici, Aldo Lipani, Yucel Saygin, and Emine Yilmaz. Evaluation Metrics for Measuring Bias in Search Engine Results. *Information Retrieval Journal*, 24(2):85–113, 2021. URL https://doi.org/10.1007/s10791-020-09386-w.

[61] Lukas Gienapp, Benno Stein, Matthias Hagen, and Martin Potthast. Estimating Topic Difficulty Using Normalized Discounted Cumulated Gain. In Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe

Cudré-Mauroux, editors, *29th ACM International Conference on Information and Knowledge Management (CIKM 2020)*, pages 2033–2036. ACM, October 2020. doi:10.1145/3340531.3412109.

[62] Martin Gleize, Eyal Shnarch, Leshem Choshen, Lena Dankin, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. 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, July 2019.

[63] Alex Graves and Juergen Schmidhuberab. 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, 2005.

[64] Nancy L. Green. Representation of Argumentation in Text with Rhetorical Structure Theory. *Argumentation*, 24:181–196, 2010.

[65] Shai Gretz, Yonatan Bilu, Edo Cohen-Karlik, and Noam Slonim. The work-week 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, November 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.findings-emnlp.47. URL `https://www.aclweb.org/anthology/2020.findings-emnlp.47`.

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

[67] Karl Gyllstrom and Marie-Francine Moens. Clash of the Typings - Finding Controversies and Children's Topics Within Queries. In Paul D. Clough, Colum Foley, Cathal Gurrin, Gareth J. F. Jones, Wessel Kraaij, Hyowon Lee, and Vanessa Murdock, editors, *Proceedings of the 33rd European Conference on IR Research (ECIR 2011)*, volume 6611 of *Lecture Notes in Computer Science*, pages 80–91. Springer, 2011. doi:10.1007/978-3-642-20161-5_10. URL `https://doi.org/10.1007/978-3-642-20161-5_10`.

[68] Ivan Habernal and Iryna Gurevych. 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. Association for Computational Linguistics, 2015. URL `http://aclweb.org/anthology/D15-1255`.

[69] Ivan Habernal and Iryna Gurevych. 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, 2016.

[70] Ivan Habernal and Iryna Gurevych. 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, 2016.

[71] Ivan Habernal and Iryna Gurevych. Argumentation Mining in User-Generated Web Discourse. *Computational Linguistics*, 2016.

[72] Ivan Habernal, Judith Eckle-Kohler, and Iryna Gurevych. Argumentation Mining on the Web from Information Seeking Perspective. In *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*, Aachen, Germany, 2014. CEUR Workshop Proceedings.

[73] Shohreh Haddadan, Elena Cabrio, and Serena Villata. 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, July 2019.

[74] Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. A Retrospective Analysis of the Fake News Challenge Stance Detection Task. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1859–1874, 2018. URL https://www.aclweb.org/anthology/C18-1158.pdf.

[75] J. A. Hartigan and M. A. Wong. A k-means clustering algorithm. *JSTOR: Applied Statistics*, 28:100–108, 1979.

[76] Kazi Saidul Hasan and Vincent Ng. 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, October 2014.

[77] Diana Hess. *Controversy in the classroom: The democratic power of discussion.* Routledge, New York, 2009.

[78] Terne Sasha Thorn Jakobsen, Maria Barrett, and Anders Søgaard. Spurious correlations in cross-topic argument mining. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 263–277. Association for Computational Linguistics, August 2021. URL https://aclanthology.org/2021.starsem-1.25.

[79] Kirsten Johnson, Di Jin, and Dan Goldwasser. Modeling of Political Discourse Framing on Twitter. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)*, pages 556–559. Association for the Advancement of Artificial Intelligence, 2017.

[80] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

[81] Kialo. Kialo. `www.kailo.com`, January 2020. Crawled in Jan, 2020.

[82] Dan Klein and Christopher D Manning. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, volume 1, pages 423–430. Association for Computational Linguistics, 2003.

[83] Lun-Wei Ku, Yu-Ting Liang, and Hsin-Hsi Chen. Question Analysis and Answer Passage Retrieval for Opinion Question Answering Systems. *Int. J. Comput. Linguistics Chin. Lang. Process.*, 13(3):307–326, 2008. URL `http://www.aclclp.org.tw/clclp/v13n3/v13n3a3.pdf`.

[84] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P. Gummadi, and Karrie Karahalios. Search Bias Quantification: Investigating Political Bias in Social Media and Web Search. *Inf. Retr. J.*, 22(1-2):188–227, 2019. doi:10.1007/s10791-018-9341-2. URL `https://doi.org/10.1007/s10791-018-9341-2`.

[85] Yuri Kuratov and Mikhail Y. Arkhipov. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. *CoRR*, abs/1905.07213, 2019. URL `http://arxiv.org/abs/1905.07213`.

[86] John Lafferty, Andrew McCallum, Fernando Pereira, et al. Conditional Random Fields: Probabilistic models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, volume 1, pages 282–289, 2001.

[87] Anne Lauscher, Goran Glavaš, and Simone Paolo Ponzetto. An argument-annotated corpus of scientific publications. In *Proceedings of the 5th Workshop on Argument Mining*, pages 40–46. Association for Computational Linguistics, November 2018. URL `https://aclanthology.org/W18-5206`.

[88] Tamar Lavee, Matan Orbach, Lili Kotlerman, Yoav Kantor, Shai Gretz, Lena Dankin, Michal Jacovi, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. 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, August 2019.

[89] John Lawrence and Chris Reed. Combining argument mining techniques. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 127–136, 2015.

[90] John Lawrence, Chris Reed, Colin Allen, Simon McAlister, and Andrew Ravenscroft. Mining Arguments From 19th Century Philosophical Texts Using Topic Based Modelling. In *Proceedings of the First Workshop on Argumentation Mining*, Baltimore, Maryland, 2014. Association for Computational Linguistics.

[91] John Lawrence, Jacky Visser, and Chris Reed. Harnessing rhetorical figures for argument mining. *Argument & Computation*, 8:289–310, 2017.

[92] Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500, August 2014. URL `https://www.aclweb.org/anthology/C14-1141`.

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

[94] Baoli Li, Yandong Liu, and Eugene Agichtein. CoCQA: Co-Training over Questions and Answers with an Application to Predicting Question Subjectivity Orientation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 937–946. Association for Computational Linguistics, 2008. URL `http://www.aclweb.org/anthology/D08-1098`.

[95] Xin Li and Dan Roth. Learning question classifiers: The role of semantic information. In *Natural Language Engineering*, volume 12, 2006.

[96] Marco Lippi and Paolo Torroni. 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, February 2016.

[97] Nitin Madnani, Michael Heilman, Joel Tetreault, and Martin Chodorow. 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*, Stroudsburg, Pennsylvania, 2012. Association for Computational Linguistics.

[98] William C. Mann and Sandra A. Thompson. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text - Interdisciplinary Journal for the Study of Discourse*, 8, 1988.

[99] Stefano Menini, Federico Nanni, Simone Paolo Ponzetto, and Sara Tonelli. Topic-Based Agreement and Disagreement in US Electoral Manifestos. In *Proceedings of the 2017 Conference of Empirical Methods in Natural Language Processing*, pages 2938–2944, 2017.

[100] Mark Meyer, Alan Barr, Haeyoung Lee, and Mathieu Desbrun. Generalized barycentric coordinates on irregular polygons. *Journal of Graphics Tools*, 7 (1):13–22, 2002.

[101] Tsvetomila Mihaylova, Georgi Karadzhov, Pepa Atanasova, Ramy Baly, Mitra Mohtarami, and Preslav Nakov. SemEval-2019 Task 8: Fact Checking in Community Question Answering Forums. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval 2019)*, pages 860–869, June 2019. URL `https://aclanthology.org/S19-2149`.

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

[103] Amita Misra, Brian Ecker, and Marilyn Walker. 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, September 2016. Association for Computational Linguistics. URL `https://aclanthology.org/W16-3636`.

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

[105] Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 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, 2007.

[106] Samaneh Moghaddam and Martin Ester. AQA: Aspect-based Opinion Question Answering. In Myra Spiliopoulou, Haixun Wang, Diane J. Cook, Jian Pei, Wei Wang, Osmar R. Zaïane, and Xindong Wu, editors, *Proceedings of the 11th International Conference on Data Mining Workshops (ICDMW 2011)*, pages 89–96. IEEE Computer Society, 2011.

doi:10.1109/ICDMW.2011.34. URL https://doi.org/10.1109/ICDMW.2011.34.

[107] Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiao-Dan Zhu, and Colin Cherry. SemEval-2016 Task 6: Detecting Stance in Tweets. In Steven Bethard, Daniel M. Cer, Marine Carpuat, David Jurgens, Preslav Nakov, and Torsten Zesch, editors, *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, pages 31–41. The Association for Computer Linguistics, 2016. doi:10.18653/v1/s16-1003. URL https://doi.org/10.18653/v1/s16-1003.

[108] Nona Naderi. Argumentation Mining in Parlimentary Discourse. In *Proceedings of the 11th International Conference of the Ontario Society for the Study of Argumentation*, pages 1–9, May 2016.

[109] Nona Naderi and Graeme Hirst. Classifying Frames at the Sentence Level in News Articles. In *Proceedings of Recent Advances in Natural Language Processing*, pages 536–542, 2017. doi:10.26615/978-954-452-049-6_070. URL https://doi.org/10.26615/978-954-452-049-6_070.

[110] Lily Ng, Anne Lauscher, Joel Tetreault, and Courtney Napoles. Creating a domain-diverse corpus for theory-based argument quality assessment. In *Proceedings of the 7th Workshop on Argument Mining*, pages 117–126, Online, December 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.argmining-1.13.

[111] Naoaki Okazaki. CRFsuite: A Fast Implementation of Conditional Random Fields (CRFs). 2007.

[112] Matan Orbach, Yonatan Bilu, Ariel Gera, Yoav Kantor, Lena Dankin, Tamar Lavee, Lili Kotlerman, Shachar Mirkin, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 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, November 2019. Association for Computational Linguistics. doi:10.18653/v1/D19-1561. URL https://www.aclweb.org/anthology/D19-1561.

[113] Matan Orbach, Yonatan Bilu, Assaf Toledo, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 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, July 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.633. URL https://www.aclweb.org/anthology/2020.acl-main.633.

[114] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report 1999-66, Stanford InfoLab, 1999. URL http://ilpubs.stanford.edu:8090/422/. Previous number = SIDL-WP-1999-0120.

[115] Raquel Mochales Palau and Marie-Francine Moens. Argumentation Mining: The Detection, Classification and Structure of Arguments in Text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 98–107. ACM, 2009.

[116] Bo Pang and Ravi Kumar. Search in the lost sense of "query": Question formulation in web search queries and its temporal changes. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 135–140, 2011.

[117] Joonsuk Park and Claire Cardie. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38. Association for Computational Linguistics, 2014. URL http://aclweb.org/anthology/W14-2105.

[118] Joonsuk Park and Claire Cardie. 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)*, May 2018.

[119] Debjit Paul, Juri Opitz, Maria Becker, Jonathan Kobbe, Graeme Hirst, and Anette Frank. Argumentative relation classification with background knowledge. In *Computational Models of Argument*, pages 319–330. IOS Press, 2020.

[120] Andreas Peldszus. Towards Segment-based Recognition of Argumentation Structure in Short Texts. In *Proceedings of the First Workshop on Argumentation Mining*, Baltimore, Maryland, 2014. Association for Computational Linguistics.

[121] Andreas Peldszus and Manfred Stede. From Argument Diagrams to Argumentation Mining in Texts: A Survey. *International Journal of Cognitive Informatics and Natural Intelligence*, 7(1):1–31, 2013.

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

[123] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 1532–1543. ACL, 2014. doi:10.3115/v1/d14-1162.

[124] Isaac Persing and Vincent Ng. End-to-End Argumentation Mining in Student Essays. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1384–1394. Association for Computational Linguistics, 2016. doi:10.18653/v1/N16-1164. URL `http://aclweb.org/anthology/N16-1164`.

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

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

[127] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *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, 2018.

[128] Martin Potthast, Lukas Gienapp, Florian Euchner, Nick Heilenkötter, Nico Weidmann, Henning Wachsmuth, Benno Stein, and Matthias Hagen. Argument Search: Assessing Argument Relevance. In *42nd International ACM Conference on Research and Development in Information Retrieval (SIGIR 2019)*. ACM, July 2019. doi:10.1145/3331184.3331327. URL `http://doi.acm.org/10.1145/3331184.3331327`.

[129] Prakash Poudyal, Jaromir Savelka, Aagje Ieven, Marie Francine Moens, Teresa Goncalves, and Paulo Quaresma. ECHR: Legal corpus for argument mining. In *Proceedings of the 7th Workshop on Argument Mining*, pages 67–75, Online, December 2020. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.argmining-1.8`.

[130] Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie L Webber. The Penn Discourse Treebank 2.0. In

Proceedings of the Sixth International Conference on Language Resources and Evaluation, 2008.

[131] Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659. Association for Computational Linguistics, August 2013. URL `https://www.aclweb.org/anthology/P13-1162`.

[132] Nils Reimers and Iryna Gurevych. Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. *arXiv preprint arXiv:1707.09861*, 2017.

[133] Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 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, 2019.

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

[135] Patrick Riehmann, Dora Kiesel, Martin Kohlhaas, and Bernd Fröhlich. Visualizing a thinker's life. *IEEE Transactions on Visualization and Computer Graphics*, 2018.

[136] Ruty Rinott, Lena Dankin, Carlos Alzate Perez, M. Mitesh Khapra, Ehud Aharoni, and Noam Slonim. 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*, pages 440–450. Association for Computational Linguistics, 2015. URL `http://aclweb.org/anthology/D15-`.

[137] Stephen Robertson and Hugo Zaragoza. *The probabilistic relevance framework: BM25 and beyond.* Now Publishers Inc, 2009.

[138] Stephen E. Robertson, Hugo Zaragoza, and Michael J. Taylor. Simple BM25 extension to multiple weighted fields. In David A. Grossman, Luis Gravano, ChengXiang Zhai, Otthein Herzog, and David A. Evans, editors,

Proceedings of the 2004 ACM CIKM International Conference on Information and Knowledge Management, Washington, DC, USA, November 8-13, 2004, pages 42–49. ACM, 2004. doi:10.1145/1031171.1031181. URL https://doi.org/10.1145/1031171.1031181.

[139] Niall Rooney, Hui Wang, and Fiona Browne. Applying Kernel Methods to Argumentation Mining. In *Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference*, 2012.

[140] Allen Roush and Arvind Balaji. Debatesum: A large-scale argument mining and summarization dataset. In *Proceedings of the 7th Workshop on Argument Mining*, pages 1–7, Online, December 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.argmining-1.1.

[141] Christos Sardianos, Ioannis Manousos Katakis, Georgios Petasis, and Vangelis Karkaletsis. Argument Extraction from News. In *Proceedings of the Second Workshop on Argumentation Mining*, Denver, Colorado, 2015. Association for Computational Linguistics.

[142] Eyal Schnarch, Carlos Alzate, Lena Dankin, Martin Gleize, Yufang Hou, Leshem Choshen, Ranit Aharonov, and Noam Slonim. 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, 2018.

[143] Claudia Schulz, Steffen Eger, Johannes Daxenberger, Tobias Kahse, and Iryna Gurevych. Multi-task learning for argumentation mining in low-resource settings. In *NAACL*, 2018.

[144] Carlos Silla and Alex Freitas. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22: 31–72, 01 2011.

[145] Maria Skeppstedt, Andreas Peldszus, and Manfred Stede. 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, 2018.

[146] Swapna Somasundaran and Janyce Wiebe. Recognizing Stances in Ideological On-Line Debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124, 2010. URL http://aclweb.org/anthology/W10-0214.

[147] Dhanya Sridhar, Lise Getoor, and Marilyn Walker. Collective stance classification of posts in online debate forums. In *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, pages 109–117, 2014.

[148] Christian Stab and Iryna Gurevych. Annotating Argument Components and Relations in Persuasive Essays. In *Proceedings of the the 25th International Conference on Computational Linguistics*, Dublin, Ireland, 2014. Dublin City University and Association for Computational Linguistics.

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

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

[151] Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. 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, June 2018.

[152] Christian Stab, Tristan Miller, and Iryna Gurevych. Cross-topic Argument Mining from Heterogeneous Sources Using Attention-based Neural Networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, page 3664–3674, 2018.

[153] Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. Cross-topic Argument Mining from Heterogeneous Sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, pages 3664–3674, Brussels, Belgium, November 2018. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/D18-1402`.

[154] Christian Matthias Edwin Stab. *Argumentative Writing Support by Means of Natural Language Processing*. PhD thesis, Technische Universität Darmstadt, 2017.

[155] Benno Stein, Yamen Ajjour, Roxanne El Baff, Khalid Al-Khatib, Philipp Cimiano, and Henning Wachsmuth. Same Side Stance Classification. In Yamen Ajjour, Khalid Al-Khatib, Philipp Cimiano, Roxanne El Baff, Basil

Ell, Benno Stein, and Henning Wachsmuth, editors, *Same Side Stance Classification Shared Task 2019*, volume 2921 of *CEUR Workshop Proceedings*, July 2021. URL `http://ceur-ws.org/Vol-2921/`.

[156] Aixin Sun and Ee-Pen Lim. 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, 2001.

[157] Maite Taboada and William C. Mann. Rhetorical Structure Theory: Looking Back and Moving Ahead. *Discourse Studies*, 8(3):423–459, 2006.

[158] Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 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, 2016. URL `https://doi.org/10.1145/2872427.2883081`.

[159] Simone Teufel. *Argumentative Zoning: Information Extraction from Scientific Text*. PhD thesis, University of Edinburgh, 1999.

[160] Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. Automatic Argument Quality Assessment–New Datasets and Methods. 2019.

[161] Orith Toledo-Ronen, Matan Orbach, Yonatan Bilu, Artem Spector, and Noam Slonim. Multilingual argument mining: Datasets and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 303–317. Association for Computational Linguistics, November 2020. URL `https://www.aclweb.org/anthology/2020.findings-emnlp.29`.

[162] Stephen Toulmin. *The Uses of Argument*. Cambridge University Press, 1958.

[163] Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, volume 1, pages 173–180. Association for Computational Linguistics, 2003.

[164] Nhat Tran and Diane Litman. Multi-task learning in argument mining for persuasive online discussions. In *Proceedings of the 8th Workshop on Argument Mining*, pages 148–153, 2021.

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

[166] Oren Tsur, Dan Calacci, and David Lazer. A Frame of Mind: Using Statistical Models for Detection of Framing and Agenda Setting Campaigns. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 1629–1638, July 2015.

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

[168] Frans H. van Eemeren and Rob Grootendorst. *A Systematic Theory of Argumentation: The Pragma-Dialectical Approach*. Cambridge University Press, Cambridge, UK, 2004.

[169] Michael Völske, Pavel Braslavski, Matthias Hagen, Galina Lezina, and Benno Stein. What Users Ask a Search Engine: Analyzing One Billion Russian Question Queries. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management (CIKM 15)*, pages 1571–1580. ACM, October 2015. URL `https://doi.org/10.1145/2806416.2806457`.

[170] Claes H. De Vreese. News Framing: Theory and typology. *Information Design Journal and Document Design*, 13(1), 2005.

[171] Henning Wachsmuth and Till Werner. Intrinsic Quality Assessment of Arguments. In *28th International Conference on Computational Linguistics (COLING 2020)*, pages 6739–6745. International Committee on Computational Linguistics, December 2020. doi:10.18653/v1/2020.coling-main.592. URL `https://aclanthology.org/2020.coling-main.592`.

[172] Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Alberdingk Tim Thijm, Graeme Hirst, and Benno Stein. Computational Argumentation Quality Assessment in Natural Language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187. Association for Computational Linguistics, 2017. URL `http://aclweb.org/anthology/E17-1017`.

[173] Henning Wachsmuth, Martin Potthast, Khalid Al-Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. Building an Argument Search Engine for the Web. In Kevin Ashley, Claire Cardie, Nancy Green, Iryna Gurevych, Ivan Habernal, Diane Litman, Georgios Petasis, Chris Reed, Noam Slonim, and Vern Walker, editors, *4th Workshop on Argument Mining (ArgMining 2017) at EMNLP*, pages 49–59. Association for Computational Linguistics, September 2017. URL `https://www.aclweb.org/anthology/W17-5106`.

[174] Henning Wachsmuth, Benno Stein, and Yamen Ajjour. ”PageRank” for Argument Relevance. In Phil Blunsom, Alexander Koller, and Mirella Lapata, editors, *15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, pages 1116–1126. Association for Computational Linguistics, April 2017. URL `http://aclweb.org/anthology/E17-1105`.

[175] Ingmar Weber, Venkata Rama Kiran Garimella, and Erik Borra. Mining Web Query Logs to Analyze Political Issues. In Noshir S. Contractor, Brian Uzzi, Michael W. Macy, and Wolfgang Nejdl, editors, *Proceedings of the 2012 Web Science Conference (WebSci 2012)*, pages 330–334. ACM, 2012. doi:10.1145/2380718.2380761. URL `https://doi.org/10.1145/2380718.2380761`.

[176] Michael Wojatzki and Torsten Zesch. ltl.uni-due at SemEval-2016 Task 6: Stance Detection in Social Media Using Stacked Classifiers. In Steven Bethard, Daniel M. Cer, Marine Carpuat, David Jurgens, Preslav Nakov, and Torsten Zesch, editors, *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, pages 428–433. The Association for Computer Linguistics, 2016. doi:10.18653/v1/s16-1069. URL `https://doi.org/10.18653/v1/s16-1069`.

[177] Eduardo Xamena, Nélida Beatriz Brignole, and Ana Gabriela Maguitman. A structural analysis of topic ontologies. *Inf. Sci.*, 421:15–29, 2017. doi:10.1016/j.ins.2017.08.081.

[178] Yunjie Xu and Zhiwei Chen. Relevance judgment: What do information users consider beyond topicality? *Journal of the American Society for Information Science and Technology*, 57(7):961–973, 2006.

[179] Taha Yasseri, Robert Sumi, András Rung, András Kornai, and János Kertész. Dynamics of conflicts in wikipedia. *PLOS ONE*, 7(6):1–12, 06 2012. URL `https://doi.org/10.1371/journal.pone.0038869`.

[180] Elad Yom-Tov, Susan Dumais, and Qi Guo. Promoting Civil Discourse through Search Engine Diversity. *Social Science Computer Review*, 32(2): 145–154, 2014.

[181] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *ACM SIGIR Forum*, volume 51, pages 268–276. ACM, 2017.

[182] Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 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)*, June 2016.