Machine Learning Driven Argument Mining

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität München

eingereicht von
Michael Fromm
am 18. Februar 2022

1. Gutachter:	Prof. Dr. Thomas Seidl
2. Gutachter:	Prof. Dr. Ralf Schenkel
3. Gutachter:	Prof. Dr. Kristian Kersting
Tag der mündlichen Prüfung:	29.07.2022

Eidesstattliche Versicherung

Hiermit erkläre ich, Michael Fromm, an Eides statt, dass die vorliegende Dissertation ohne unerlaubte Hilfe gemäß Promotionsordnung vom 12.07.2011, §8, Abs. 2 Pkt. 5, angefertigt worden ist.

München, 18. Februar 2022 _____

Michael Fromm

Acknowledgements

I want to thank several people for their tremendous support while writing this thesis. First, I would like to acknowledge my external reviewers, Prof. Dr. Kristian Kersting and Prof. Dr. Ralf Schenkel, for their willingness to review my thesis and the interesting debate during the defense.
I would be remiss in not mentioning all the students I supervised in their thesis and practical courses. Thank you for your support!
Thanks should also go to my colleagues and collaborators who helped me develop these thesis ideas. In particular, I would like to thank the following colleagues:
I sincerely thank Prof. Dr. Matthias Schubert for his support and guidance.
I'm extremely grateful to my supervisor Prof. Dr. Thomas Seidl, who always gave me the freedom in research to pursue my ideas. Thank you for your encouragement and trust!
I also want to thank Max Berrendorf for the many discussions about research ideas, and the late-night feedback and editing sessions! Thank you for your help; I learned a lot from you!
I am incredibly grateful to Evgeniy Faerman, who introduced me to Machine Learning, especially neural networks. He guided me through the first years of my Ph.D. study. Thank you for your unwavering support and belief in me!
I want to express my deepest gratitude to my family for their help and reassurance. I could not have undertaken this journey without the aid of my parents, Andreas and Monika, who gave me the freedom to follow my dreams. I also would like to thank my son Robin, who joined us when I was writing my dissertation, for giving me unlimited happiness and pleasure. Finally, this endeavor would not have been possible without my wife, Andrea. She moved to Munich with me and supported me even before I started my university education. Thank you for having patience with me and always believing in me!

Contents

Abstract

For thousands of years, humans have utilized argumentation to achieve a holistic understanding by presenting and contesting different views. Nowadays, achieving a common understanding in society becomes more challenging since disputes are often global, involve various interest groups, and the amount of relevant and often contradicting arguments grows correspondingly. On top of that, discussions no longer occur only on talk shows and at family tables but also online. As a result, relevant arguments can be found on online platforms such as social media, debate portals, or as comments on news portals. Therefore manual identification, retrieval, or even comparison of already extracted arguments becomes infeasible. This thesis focuses on solving these problems automatically by utilizing Machine Learning. The main goal is to enable methods to learn the concepts of argumentation automatically, without providing explicit rules and definitions of arguments and their relationships.

We start with the problem of argument identification in heterogeneous text sources. The objective is to extract supporting and opposing arguments for highly controversial topics from various text sources, which can be found online. We identify that the relation between the topic and the arguments is crucial and investigate different possibilities to incorporate the topic information in the identification process.

Furthermore, we focus on argument identification in peer reviews for scientific publications. We demonstrate that arguments in peer reviews have their peculiarities and that knowledge transfer from other domains is only possible to a limited extent. Therefore we provide the community with a newly developed peer-review dataset from multiple conferences. Our work shows that peer reviews contain a broad range of arguments, and these arguments can also be precisely and automatically identified with a suitable model. In addition, we highlight that arguments drive the peer-reviewing process in research and that these arguments are decisive for the publication decision.

Next, we address the problem of argument retrieval. Even the most comprehensive collection of arguments is only helpful if the suitable arguments can be retrieved at the right time. The crucial challenge is identifying the relevant documents, covering all relevant aspects, and not repeating themselves. We demonstrate how Machine Learning methods can be helpful for this task and that the proper selection and design of the training task plays a crucial role for the performance.

Lastly, we investigate the decisive argument property of strength or quality. Solid arguments help to convince others, to compromise, or provide for a better understanding. While other work often considers argument quality in isolation, we link it with further Argument Mining tasks, assess generalization across various text domains, study the impact of emotions, and evaluate the impact of assumptions made in the previous studies.

Zusammenfassung

Seit Tausenden von Jahren nutzen Menschen Argumente, um ein ganzheitliches Verständnis zu erreichen, indem unterschiedliche Ansichten dargelegt und bestritten werden. Allerdings wird es heutzutage immer schwieriger, ein gemeinsames Verständnis in der Gesellschaft zu erreichen. Debatten sind mittlerweile häufig global, involvieren verschiedene Interessengruppen und die Menge an relevanten und oft widersprüchlichen Argumenten wächst entsprechend. Hinzu kommt, dass Diskussionen nicht mehr nur in Talkshows und am Esstisch stattfinden, sondern auch online, und dass relevante Argumente auch auf Online-Plattformen wie Social Media, Debattenportalen oder als Kommentare auf Nachrichtenportalen zu finden sind. Daher ist eine manuelle Identifizierung, das Auffinden und der Vergleich bereits extrahierter Argumente nicht mehr praktikabel. Diese Arbeit konzentriert sich auf die automatische Lösung dieser Probleme mit Hilfe von maschinellem Lernen. Das Hauptziel ist, Verfahren zu entwickeln, welche die Konzepte der Argumentation automatisch erlernen, ohne explizite Regeln und Definitionen von Argumenten und deren Beziehungen zu liefern.

Wir beginnen mit dem Problem der Identifizierung von Argumenten in heterogenen Textquellen. Hier sollen unterstützende und gegensätzliche Argumente für hochkontroverse Themen aus verschiedenen Textquellen extrahiert werden. Wir stellen fest, dass die Beziehung zwischen Thema und Argumenten von entscheidender Bedeutung ist. Es werden verschiedene Methoden empirisch untersucht, welche die Themeninformation in den Identifikationsprozess einbeziehen.

Außerdem konzentrieren wir uns auf die Identifizierung von Argumenten in Peer-Reviews für wissenschaftliche Publikationen. Wir zeigen, dass Argumente in Peer-Reviews besondere Merkmale haben und dass ein Transfer aus anderen Domänen nur bedingt möglich ist. Deshalb entwickeln wir einen neuen Datensatz mit Peer-Reviews von mehreren Konferenzen und stellen ihn der Allgemeinheit zur Verfügung. Unsere Arbeit zeigt, dass Peer-Reviews ein breites Spektrum an Argumenten enthalten, und dass diese Argumente mit einem geeigneten Verfahren präzise und automatisch identifiziert werden können. Darüber hinaus legen wir dar, dass Argumente für den Peer-Review-Prozess und die Publikationsentscheidung essentiell sind.

Schließlich gehen wir auf das Problem der Suche nach Argumenten ein. Selbst die umfangreichste Sammlung von Argumenten ist nur dann hilfreich, wenn die passenden Argumente zum richtigen Zeitpunkt abgerufen werden können. Die entscheidende Herausforderung besteht darin, die relevanten Argumente zu identifizieren, welche alle ausschlaggebenden Aspekte abdecken ohne sich dabei zu wiederholen. Wir zeigen, wie Methoden des maschinellen Lernens bei dieser Aufgabe hilfreich sein können, und dass die richtige Auswahl und Gestaltung des Trainings eine entscheidende Rolle für die Leistung der Argument-Suchmaschine spielt.

Abschließend untersuchen wir die Stärke und Qualität von Argumenten. Solide Argumente helfen dabei, andere zu überzeugen, Kompromisse zu schließen oder schlichtweg für ein besseres Verständnis zu sorgen. Während andere Arbeiten die Qualität von Argumenten oft isoliert betrachten, kombinieren wir diese mit weiteren Problemstellungen des Argument Mining, bewerten die Übertragbarkeit über verschiedene Textdomänen hinweg, untersuchen den Einfluss von Emotionen und bewerten die Auswirkungen der in den vorherigen Studien getroffenen Annahmen.

List of Publications and Declaration of Authorship

Chapter 2

The Chapter 2 corresponds to the following publication:

> **Michael Fromm**, Evgeniy Faerman, and Thomas Seidl. „TACAM: Topic And Context Aware Argument Mining.“ In: *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE. 2019, pp. 99–106

Michael Fromm proposed the research idea, developed and conceptualized it with Evgeniy Faerman, and discussed it with Thomas Seidl. Michael Fromm did the implementation, the design of the architecture and framework. Michael Fromm designed and conducted the experiments and analyzed their results. Michael Fromm and Evgeniy Faerman discussed the results and wrote the manuscript. All authors revised the manuscript.

Chapter 3

The Chapter 3 corresponds to the following publication:

> **Michael Fromm**, Evgeniy Faerman, Max Berrendorf, Siddharth Bhargava, Ruoxia Qi, Yao Zhang, Lukas Dennert, Sophia Selle, Yang Mao, and Thomas Seidl. „Argument Mining Driven Analysis of Peer-Reviews.“ In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35 (May 2021), pp. 4758–4766. URL: `https://ojs.aaai.org/index.php/AAAI/article/view/16607`

Michael Fromm, Evgeniy Faerman, and Max Berrendorf developed and conceptualized the research idea. Siddharth Bhargava, Ruoxia Qi, Yao Zhang, Lukas Dennert, Sophia Selle, and Yang Mao implemented and evaluated the approach in the practical project. Michael Fromm, Evgeniy Faerman, and Max Berrendorf analyzed the results and discussed the findings. Michael Fromm, Evgeniy Faerman and Max Berrendorf wrote the manuscript. All authors revised the manuscript.

Chapter 4

The Chapter 4 corresponds to the following publication:

> Dietrich Trautmann, **Michael Fromm**, Volker Tresp, Thomas Seidl, and Hinrich Schütze. „Relational and Fine-Grained Argument Mining.“ In: *Datenbank-Spektrum* (2020), pp. 1–7

The survey of our work was proposed, developed and conceptualized by Michael Fromm and Dietrich Trautmann. Michael Fromm, and Dietrich Trautmann wrote the manuscript and discussed it with Volker Tresp, Thomas Seidl, and Hinrich Schütze. All authors revised the manuscript.

Chapter 5

The Chapter 5 corresponds to the following publication:

> **Michael Fromm***, Max Berrendorf*, Sandra Obermeier, Thomas Seidl, and Evgeniy Faerman. „Diversity Aware Relevance Learning for Argument Search.“ In: *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part II*. vol. 12657. Lecture Notes in Computer Science. * equal contribution. Springer, 2021, pp. 264–271. DOI: 10.1007/978-3-030-72240-1_24. URL: https://doi.org/10.1007/978-3-030-72240-1_24

The research idea was proposed by Michael Fromm, developed and conceptualized by Michael Fromm and other co-authors. Michael Fromm, Max Berrendorf, and Sandra Obermeier did the implementation and the design of the architecture, the framework and the pipeline. Michael Fromm and Max Berrendorf conducted the experiments and analyzed their results. The findings were discussed with all authors. Michael Fromm, Evgeniy Faerman, Max Berrendorf, and Sandra Obermeier wrote the manuscript. All authors revised the manuscript.

Chapter 6

The Chapter 6 corresponds to the following publication:

> Nataliia Kees, **Michael Fromm**, Evgeniy Faerman, and Thomas Seidl. „Active Learning for Argument Strength Estimation.“ In: *Proceedings of the Second Workshop on Insights from Negative Results in NLP*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 144–150. URL: https://aclanthology.org/2021.insights-1.20

The research idea was proposed by Michael Fromm, developed and conceptualized by Michael Fromm and Evgeniy Faerman. Nataliia Kees did the implementation of the architecture. Michael Fromm, Nataliia Kees and Evgeniy Faerman designed and conducted the experiments. All authors discussed the findings. Michael Fromm and Nataliia Kees wrote the manuscript.

Chapter 7

The Chapter 7 corresponds to the following submitted work, that is currently under peer-review:

> **Michael Fromm**, Max Berrendorf, Johanna Reiml, Isabelle Mayerhofer, Siddharth Bhargava, Evgeniy Faerman, and Thomas Seidl. „Towards a Holistic View on Argument Quality Prediction.“ In: Peer-Reviewing Phase

The research idea was proposed, developed and conceptualized by Michael Fromm. Michael Fromm did the design of the architecture and the pipeline. Johanna Reiml, Isabelle Mayerhofer, Siddharth Bhargava did the implementation and conducted the experiments. Michael Fromm, Evgeniy Faerman, and Max Berrendorf analyzed and discussed the results. Michael Fromm, Evgeniy Faerman, and Max Berrendorf wrote the manuscript. All authors revised the manuscript.

1 Introduction

Arguments have been studied since the ancient Greeks [142] and are an essential mechanism for the development of society. In particular, acute crises such as the corona crisis show that policy and society rely on arguments from different research areas that provide evidence for decisions. Understanding, visualizing, and structuring argumentation is a key to reducing the impact of disinformation campaigns, fake news, and populism. Argumentation is multi-disciplinary and spans diverse research areas such as philosophy, logic, rhetoric, language, law, psychology, and computer science. Multiple domains on the Web such as online newspapers, blogs, or social media provide an ever-growing flow of information where arguments can be identified, ranked, and analyzed.

The availability of primarily unstructured data sources combined with recent advances in Computational Linguistics and Machine Learning (ML) provide a solid basis for *Argument Mining (AM)*. AM is a young field in Natural Language Processing (NLP) that focuses on automatically identifying and extracting the structure of arguments in texts. In contrast to other areas such as *sentiment analysis* or *opinion mining*, that focus on *what* opinions are expressed, AM aims to determine *why* someone has a particular stance regarding a specific viewpoint.

AM tasks like *relational Argument Identification (AId)* share analogies with other ML tasks such as link prediction [55] or textual entailment [126]. Thus the AM tasks can often be directly formulated as ML problems. AM relies on many ML components, such as information extraction, (world) knowledge representation, and discourse analysis. The critical element of these components are the features and the incorporation of context information.

AM applies to multiple domains. For instance, it provides benefits and value in the law domain. Legal texts are analyzed on successful and unsuccessful patterns of arguments in U.S. judicial decisions [192], arguments can be extracted from decisions of the European Court of Human Rights (ECHR) [135] or be used as a tool for case summarizations [183].

Another exciting application domain of AM is all forms of dialogues. Debates often occur in society and politics and are unique due to their vocal and paralinguistic features. AM can extract argument components (e.g., claims) in political debates [99, 71, 23, 39, 173], it can summarize the arguments in controversies [149, 156, 6] and also find attacking or supporting relations between the actors in an exchange [23, 173].

Another recent success was the IBM Project Debater [145], the first ML system to debate humans on controversial topics. It was competitive against a world champion debater in a live debate by performing a speech composed of mined arguments. It was also able to attack the arguments raised by the challenger and perform a rebuttal.

As a concrete example, consider the peer-reviews from an online conference. Sentiment analysis can give an impression of whether the reviews are *positive or negative* about

a submitted paper, AM can extract *why* they have these sentiments. An AM pipeline could consist of three steps: First, *AId* extracts the supporting and opposing arguments from the peer-reviews. In the next step, *Argument Retrieval (AR)* can eliminate vast amounts of duplicate semantic arguments and cluster the remaining arguments based on different criteria (such as novelty, the presentation, or the lack of related work). In the last step, *Argument Quality (AQ)* can provide a feedback mechanism that weights the leftover arguments based on the frequency or quality. The AM pipeline would then deliver a compact overview that reduces the time of the manual inspection of the reviews. This outline can then be used along the peer-reviews for the acceptance decision.

A common challenge in AM is that the *context* in which a text span appears can specify whether it is an argument or not. Multiple works showed that surrounding context can be more important than the content of the text span [49, 124, 120]. This is problematic for the general application of ML approaches that do not include contextual information. The context is often captured differently in some domains, and models trained in one domain can struggle to classify text spans in another domain.

Another hurdle in AM is the lack of consistently annotated data across text domains. Recent work has concentrated on generating annotations of specific text domains, e.g. persuasive essays [168, 185, 121, 129, 119, 25, 24, 167, 21] or tweets [159, 38, 16, 81, 160, 80, 117, 161]. Annotating data for these domains often consist of specific annotation schemes aimed at their respective application area, which dramatically limits generalization. Recent work adapts the complex annotation schemes and introduces more practical schemes that can be better applied to heterogeneous text domains [170, 166, 178, 49, 42].

Alongside the developments in AM, also the underlying architectures improved. Feature extraction, the most crucial preprocessing step, could be integrated inside the ML architectures by using Transfer Learning. In a first step, word embeddings [112, 128] are trained on an unsupervised task where large amounts of data are available. Afterward, they are used to initialize the first layers of architectures such as the Transformer [181] or the Long-Short-Term Memory (LSTM) [74], which is then trained on AM tasks. Recently, Transfer Learning in NLP developed even further through large language models [33]. Language models initialize whole architectures with weights that are pretrained on diverse unsupervised tasks such as masked-language modeling (MLM) and next sentence prediction (NSP) [33]. With their inherent context knowledge, these language models provide a solid basis for architectures, frameworks and pipelines in AM. [49, 50, 64].

1.1 Overview

This chapter provides a broad overview of the research area of *Argument Mining (AM)*. First, we discuss text representations and a plethora of Machine Learning (ML) architectures (see Section 1.2). Next, we discuss Transfer Learning methods and language models based on the Transformer architecture [181, 33, 78, 130], an inflection point for AM and Natural Language Processing (NLP) (see Section 1.3). These advancements allowed the use of Transfer Learning [176] broadly across different tasks in AM. Lastly, Section 1.4 establishes *Argument Identification (AId)* (see Section 1.4.2), *Argument Retrieval (AR)* (see Section 1.4.3), and *Argument Quality (AQ)* (see Section 1.4.4) as AM tasks and highlights difficulties and commonalities in each of them. Finally we position the contributions of our research papers inside the field.

1.2 Text Representation & Processing Architectures

Natural language is a rich source of information for many use-cases. The peculiarities are, e.g., their discrete and sparse space, making it a challenging data source. AM tasks require transforming text into a machine-understandable representation before any processing. The transformation is called *feature extraction* or *feature representation* and is done by a *feature function*. The representations can be created for any text unit, like subwords, words, phrases, sentences, and whole documents. These representations, usually a vector of texts features, significantly impact the AM tasks. Lippi et al. [100] even state that "the key element for achieving good performance (in AM) has shown to be the choice of the features, rather than the ML algorithm.". This section surveys and presents multiple classical and neural models that are commonly used in AM. We categorize the representation and architectures into two general classes (compare Fig. 1.1): classical and deep learning models. Classical models were mainly used at the beginning of the AM era (the first works were published early in the last decade), and the feature extractor was handcrafted for each task and dataset. Shallow neural models are a step in-between and have the advantage that the feature extractor was trained on unsupervised data and can be used across different AM tasks and architectures (more in Section 1.2.2.2). Section 1.2.3 presents deep neural networks with an *inductive bias on sequential data* such as Recurrent Neural Network (RNN) [152]. In Section 1.2.4, we lift the sequential data bias and present the attention-based Transformer architecture [181]. RNNs with word embeddings were state-of-the-art in most AM tasks until contextual embeddings (see Section 1.3).

1.2.1 Classical Models

Classical ML models were common in the dawn of AM. Models such as the Support-vector machines (SVM) [28], Random Forests [18], Decision Trees [136], Multinomial Logistic Regression [89], and Naive Bayes [82] required "general" and "task-specific" features that were carefully designed for each task. General representation techniques such as

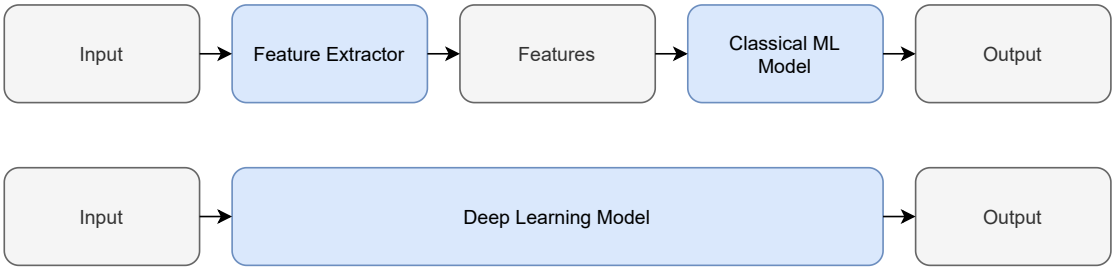

Figure 1.1: A comparison of the classical Machine Learning and the Deep Learning procedure.

bag-of-words models, One-hot Encoding, and term frequency-inverse document frequency (TF-IDF) were developed to represent words and sequences. For example, the sentence "Nuclear energy should be abolished." would be represented as follows:

1. One-hot Encoding: [nuclear → [1 0 0 0 0], energy → [0 1 0 0 0], should → [0 0 1 0 0], be → [0 0 0 1 0], abolished → [0 0 0 0 1]]
2. TF-IDF: [(nuclear, 1.23), (energy, 1.9), (should, 1.2), (be, 1.3), (abolished, 1.5)], the real numbers are exemplary for the assigned weight

The representations can be used for various tasks within a large corpus of sentences. One possible use case is predicting the next word in a sentence. However, these simple discrete representations come with many disadvantages:

1. The representation space is directly proportional with the vocabulary size
2. The feature space is sparse, with only a few non zero values
3. The representation does not capture the semantics of the word (e.g., spooky & scary should be similar to each other)
4. Assumes independence of words and does not include the context
5. The positional information is not captured, and the weights are often dependent on the domain of the corpus

Researchers in NLP and AM developed further "task-specific" features to compensate for the previously stated disadvantages. Ablation studies highlighted what features were more or less critical for the specific task. The number of advanced features across all AM papers would be too much to present and discuss here. Therefore some of the used features in AM are highlighted in Table 1.1.

These "handcrafted" features have the advantage that they have a **real-world interpretation** and therefore add explain-ability, contrary to **learned representations** (see Section 1.2.2.2) which often have less interpretation capability.

Group	Feature	Explanation
Lexical	Unigram	Binary and lemmatized unigrams
	Dependency tuples	Lemmatized dependency tuples
Structural	Token statistics	Number of tokens, paragraphs, sentences
	Component position	Rel. pos. in section, number of proc. tokens
Indicators	1st-person indicators	"I", "me", "my", present in preceding tokens
Contextual	Shared phrases	Shared noun phrases in diff. parts of the section
Syntactic	Tense of main verb	Tense of the main verb in the sentence
Probability	Type probability	Cond. prob. of the sentence being a certain class

Table 1.1: Exemplary features of an classification model

1.2.2 Shallow Neural Models

In this section, we study a multitude of different architectures for representation learning in text. We start with the Multilayer Perceptron (MLP) and investigate its ability for text processing.

1.2.2.1 Multilayer Perceptron

The most straightforward neural network is the *perceptron*. It is a simple linear model:

$$NN_{Perceptron}(x) = xW + b \tag{1.1}$$

$$x \in \mathbb{R}^{d_{in}}, W \in \mathbb{R}^{d_{in} \times d_{out}}, b \in R^{d_{out}}$$

where x is the input, W is the weight matrix, and b is the bias term.

In order to get to an MLP, we introduce a nonlinear hidden layer.

$$NN_{MLP}(x) = g(xW^1 + b^1)W^2 + b^2 \tag{1.2}$$

$$x \in \mathbb{R}^{d_{in}}, W^1 \in \mathbb{R}^{d_{in} \times d_1}, b^1 \in R^{d_1}, W^2 \in \mathbb{R}^{d_1 \times d_2}, b^2 \in R^{d_2}$$

Here W^1 and b^1 are a matrix and a bias term for the linear transformation of the input, g is an nonlinearity (e.g. a sigmoid function or the rectified linear unit rectified linear unit (ReLU) [1]) which is applied element-wise, and W^2 and b^2 are the terms of a second linear transformation. Vectors from linear transformations are referred to as *layers*. Layers resulting from linear transformations are also often referred to as *fully connected layers*.

In terms of representation power, it was shown by Hornik et al. [76] and Cybenko [29] that NN_{MLP} is a universal approximater - it can approximate with any non-zero amount of error all continuous functions on a closed and bounded subset of R^n, and all functions mapping from any finite dimensional discrete space to another.

In order to train a neural network, the procedure is similar to training a linear classifier; a loss function needs to be specified. The loss function states the loss of predicting

$\hat{y}$ when the true output is y. The most common used loss function in AM tasks such as AId (see Section 1.4.2) is *Categorical Cross-Entropy (CCE)* (see Eq. (1.3)) loss for multi-class classification (e.g. supporting-argumentative, opposing-argumentative or non-argumentative) or *Mean squared error (MSE)* (see Eq. (1.4)) for regression problems such as absolute AQ estimation (see Section 1.4.4).

The CCE Loss is defined as follows:

$$CCE = -\sum_{i}^{C} t_i(log\Big(\frac{e^{x_i}}{\sum_{j}^{C} e^{x_j}}\Big) \tag{1.3}$$

where C is the number of classes, t_i is the target vector where only one dimension is one, and x are the the neural network output vectors.

The MSE is defined as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1.4}$$

where n predictions are generated from a sample of n data points.

MLPs have been used relatively infrequently (compared to e.g., SVM and other "classical" approaches) in AM. One reason is their worse performance on small datasets, which were common in the early period of AM. However, MLPs are often used for the learning of word embeddings (see Section Section 1.2.2.2) and also build the foundation for RNNs and the Transformer architecture.

1.2.2.2 Word Embeddings

Unlike sparse- and independent representations such as One-hot encoding (see Section 1.2.1), *word embeddings* are densely distributed and dependent on other (nearby) words. Word embeddings represent various metrics and concepts (e.g., context) in the embedding dimensions. The information about a word is distributed along the vector dimensions. Consequently, the embeddings are called *distributed text representations*. The embeddings can be generated by methods such as neural networks [112], word co-occurrence matrices [90, 93, 96], probabilistic models [59], knowledge base methods [137], or by explicit representation in contextual terms [92]. In this thesis, we will present the two most common word embeddings: Word2Vec [110, 111] and Global Vectors (GloVe) [128]. Both are common in AM and are generated by neural networks. Word2Vec is a predictive embedding model that can use two types of architectures to generate vector representations of words. *Continuous Bag-of-Words (CBOW)* and Skip-grams. Both rely on MLP and are trained using stochastic gradient descent and backpropagation [152].

1.2.2.3 Continuous Bag-of-Words

CBOW predicts the most likely word w in the given context c (see Fig. 1.2). Words with a similar likelihood of appearing in a particular context c are considered similar and therefore closer in the vector-space . Given the text corpus T, the goal is to optimize the parameter Θ of the conditional probability $p(w|c;\Theta)$ to maximize the corpus probability as follows:

$$\arg\max_{\Theta} \prod_{w \in T} \Big[\prod_{c \in C(w)} p(w|c;\Theta) \Big] \tag{1.5}$$

where $C(w)$ is the set of contexts of word w.

1.2.2.4 Skip-Gram

Skip-gram predicts the most likely context c, using a given word w (see Fig. 1.2). Given the text corpus T, the goal is to optimize the parameter Θ of the conditional probability $p(c|w;\Theta)$ to maximize the corpus probability as follows:

$$\arg\max_{\Theta} \prod_{w \in T} \Big[\prod_{c \in C(w)} p(c|w;\Theta) \Big] \tag{1.6}$$

where $C(w)$ is the set of contexts of word w.

In a large corpus with a vast number of word embedding dimensions, the *skip-gram model* yields the highest performance in terms of overall accuracy [110]. However, *CBOW* is less computationally expensive (and therefore several times faster to train) and yields similar results [110] as the skip-gram model.

1.2.2.5 Global Vectors

GloVe [128] are based not only on local statistics of the corpus such as Word2Vec (context window) but also incorporating global corpus statistics in the form of word co-occurrences to obtain word vectors. Using global corpus statistics to gain semantic information dates back to the latent semantic analysis (LSA) [34]. GloVe and LSA hypothesize that the ratios of word-word co-occurrences encode a form of meaning in the representations. Table 1.2 highlights an example with probabilities from a six billion word corpus. One can see that *ice* co-occurs more frequently with *solid* than with *gas*, and *steam* co-occurs more frequently with *gas* than with *solid*. The ratio of probabilities cancels out noise from non-discriminative words like *water* and *fashion*.

The example above shows that a starting point for word vector learning should be the ratios of co-occurrence probabilities rather than probabilities themselves. The most general form of the model is:

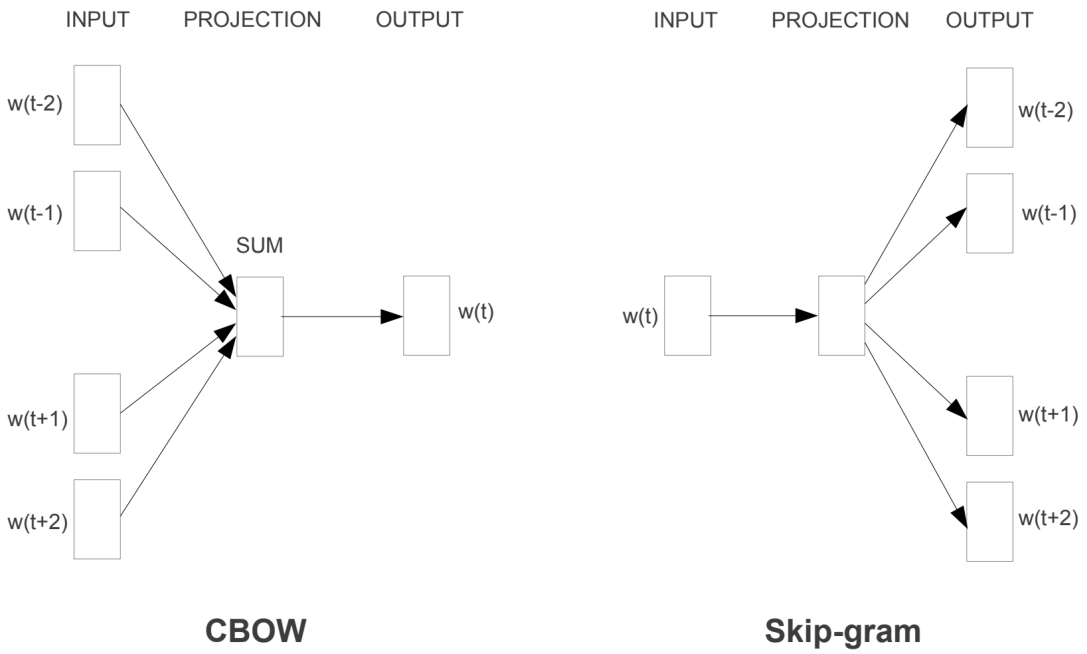

Figure 1.2: The Skip-gram predicts context words given the *current word* and the Continuous Bag-of-Words model predicts the current word based on the *context*. Original figure from [110].

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

Table 1.2: Example of the co-occurrence probabilities and ratio of the target words *ice* and *steam* with various words from the vocabulary [128].

$$F(w_i, w_j, w_k) = \frac{P(k|i)}{P(k|j)} \tag{1.7}$$

where F is a learned function, $w_{i,j,k} \in \mathbb{R}^d$ are word vectors and $\frac{P(k|i)}{P(k|j)}$ is extracted from the corpus. Casting Eq. (1.7) in a least-square function with a vocabulary V yields:

$$J = \sum_{i,j}^{V} f(X_{i,j}) \Big[w_i^T w_j + b_i + b_j - \log(X_{i,j)} \Big]^2 \tag{1.8}$$

where f is a weighting function [128], $X_{i,j}$ is the cell i, j of the word-word co-occurrence matrix, V is the size of the vocabulary, w_i, w_j are the vector representation of word i, and the context j, and b their corresponding bias term. The function f is required as the $\log(X_{i,j})$ diverges whenever its argument is zero. The derivation is presented more detailed in Pennington et al. [128].

GloVe embeddings achieve superior results compared to Word2Vec embedding results on the *word analogy* task [112] such as "a is to b as c is to _", on a variety of *word similarity* tasks [45, 113, 151, 79, 103], and also in *named entity recognition* (NER) tasks [157, 109].

A general problem with word embedding methods is that they do not consider *polysemy* and *homonymy*. A word is polysemous if it encodes different but related meanings. For example, the *newspaper* is both a company and a piece of paper in the following two sentences.

- The enraged men sued the newspaper.
- The newspaper ignited the fire.

Homonymy occurs if two unrelated words look or sound similar, as the word *bark* in the following two sentences.

- The dogs bark at the neighbor.
- The bark of the tree is light brown.

Since word vectors represent the average of the contexts they appear in, they will not always represent the type of similarity that we are after. Also, the definition of similarity can be a problem since some facets of similarity can be harmful to specific tasks. For example, *London* and *Berlin* might be close to each other in the embedding space as both are capitals. However, for a flight booking system their dissimilarity however would be significant. Another problem is the *lack of context*. The distributional approach aggregates the context in which a word appears in a large corpus. This results in *context-independent* representations. In reality, however, there is no such entity as a context-independent

word as argued by Firth [46], *"the complete meaning of a word is always contextual, and no study of meaning apart from context can be taken seriously"*. Deep neural models presented in the two subsequent Section 1.2.3 and Section 1.2.4 attempt to lessen the problem of polysemy and homonymy and the definition of similarity by looking at larger text structures such as whole sentences, paragraphs, or whole documents.

1.2.3 Recurrent Neural Networks

This subsection introduces neural architectures, called *Recurrent Neural Network (RNN)s* with a strong *inductive bias* for language data or sequences. We further look at derived architectures such as the Long Short-Term Memory (LSTM) [74], the Contextual Long Short-Term Memory (CLSTM) [56], and the attention-based LSTM [4]. RNNs are designed to process sequences or text units, detect patterns and regularities, and allow the model to investigate dependencies in the sequence. An RNN is not a standalone component but rather a *feature extractor* that produces a vector that is fed into a different part of the network, e.g., a *Multilayer Perceptron (MLP) classifier*. Most RNNs are trained *end-to-end*, which means that the feature extractor is trained to be helpful for the given prediction task. RNNs allow the representation of arbitrarily sized sequences while paying attention to the order of the inputs.

The simplest RNN is the Elman Network or Simple-RNN (S-RNN) [43]. The state s_i of an S-RNN is defined as follows:

$$s_i(x_i, s_{i-1}) = g(s_{i-1}W_s + x_iW_x + b) \tag{1.9}$$

$$s_i, y_i \in \mathbb{R}^{d_s}, x_i \in \mathbb{R}^{d_x}, W_x \in \mathbb{R}^{d_x \times d_s}, W_s \in \mathbb{R}^{d_s \times d_s}, b \in \mathbb{R}^{d_s}$$

here x_i denotes the input vector x, at the position i, of a sequence of length n, the other input is the previous state s_{i-1} of the RNN. The weight matrices linearly transform both inputs, and the results are summed up (together with a bias term) and then passed through a nonlinear activation function g (most of the times a sigmoid or rectified linear unit (ReLU)). The output y_i is equal to the hidden state s_i at position i, and can be used for further predictions. The S-RNN is hard to train effectively because of the vanishing and exploding gradients [127]. The error signal for later elements of the sequence vanishes in the backpropagation process, making it hard to capture and learn long-range dependencies.

1.2.3.1 Bidirectional RNN

Unidirectional RNN only preserves past information since the state s_i only depends on x_i and s_{i-1}. Bidirectional RNN are implemented as two unidirectional RNNs, one regularly from the past to the future and the other backwards from the future to the past. Layers are then aggregated (e.g., by concatenation or summarizing) of the two states:

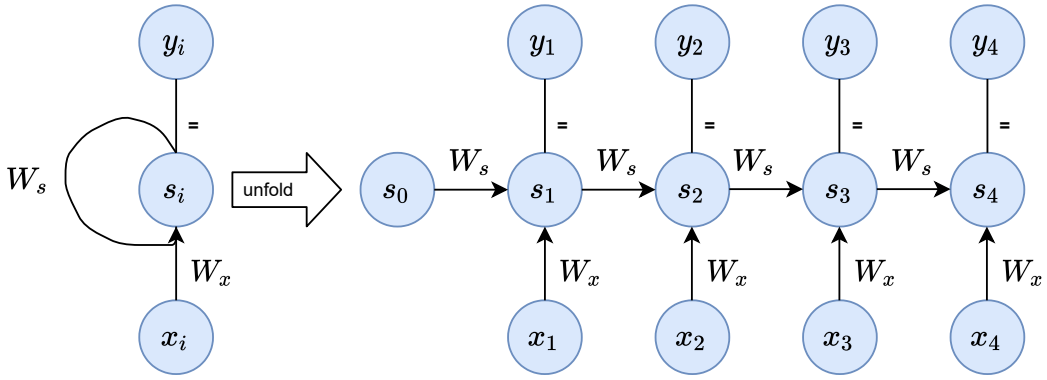

Figure 1.3: Unrolling of a Simple-RNN (S-RNN) [43]

$$s_i = aggr(s_i^{\rightarrow}, s_i^{\leftarrow}) \tag{1.10}$$

where the arrows indicate the direction of the RNN.

Bidirectional Long Short-Term Memory (BiLSTM) (see the next Section 1.2.3.2) show very good results on Argument Mining (AM) tasks [170, 41, 49] as they can capture the context better than their unidirectional counterpart.

1.2.3.2 Long Short-Term Memory

The LSTM architecture [74] was developed to solve the vanishing and exploding gradients problem of RNNs. In the Simple-RNN architecture (see the previous Section 1.2.3), the repeated multiplications of the weight matrices W_x and W_s make it very likely for the values to explode or vanish. The LSTM is the first architecture that constructs a gating mechanism. Instead of a single state vector s, the architecture uses a *memory cell* representation c and a *working memory* h comparable to the state s of the Simple-RNN. Multiple gates are responsible for deciding which part of the new input should be *written* to the memory state and what should be *deleted*. The state s_j of a LSTM is defined as[1]:

$$\begin{aligned} s_j(x_j, s_{j-1}) &= [c_j; h_j] \\ c_j &= f \odot c_{j-1} + i \odot z \\ y_j = h_j &= o \odot tanh(c_j) \\ i &= \sigma(x_j W_{xi} + h_{j-1} W_{hi}) \\ f &= \sigma(x_j W_{xf} + h_{j-1} W_{hf}) \end{aligned} \tag{1.11}$$

[1]There exist a huge variety of LSTM architectures (e.g. with different gates), the architecture presented here, is the most common used. For an overview and empirical comparison of different architectures, see Greff et al. [63]

$$o = \sigma(x_j W_{xo} + h_{j-1} W_{ho})$$
$$z = tanh(x_j W_{xi} + h_{j-1} W_{hz})$$
$$s_j \in \mathbb{R}^{2d_h}, x_i \in \mathbb{R}^{d_x}, c_j, h_j, i, f, o, z \in \mathbb{R}^{d_h}, W_{x\_} \in \mathbb{R}^{d_x \times d_h}, W_{h\_} \in \mathbb{R}^{d_h \times d_h}$$

The state of the LSTM at timestep j is compiled of c_j, the memory component, and h_j, the hidden state. Usually, word embeddings are used as the input representation of x (see previous Section 1.2.2.2). The three gates, **i**, **f**, and **o**, control the **i**nput, the **f**orgetting, and the **o**utput of the input x_j and the previous state h_{j-1}. σ and $tanh$ indicate the activation function. The memory c_j is updated by forgetting ($f \odot c_{j-1}$) parts of the old memory state c_{j-1} and writing ($i \odot z$) parts of the new input x_j. The output gate o controls what can be used from the current memory state c_j. The output vector y_j can then be used for further predictions.
LSTM have been widely used in the area of Argument Identification (AId) (see Section 1.4.2) [134, 16, 41, 65, 147, 170, 49], and also in Argument Quality (AQ) estimation (see Section 1.4.4) [69, 70, 64].

1.2.3.3 Contextual LSTM

A special LSTM variant is the CLSTM [56]. The model allows the incorporation of additional context (e.g., a topic in the AId setting). The context can improve the performance compared to the standard LSTM model [56, 170]. In the following equations, the term in bold is the modification to the original LSTM equation.

$$i = \sigma(x_j W_{xi} + h_{j-1} W_{hi} + \boldsymbol{t_j W_T})$$
$$f = \sigma(x_j W_{xf} + h_{j-1} W_{hf} + \boldsymbol{t_j W_T})$$
$$o = \sigma(x_j W_{xo} + h_{j-1} W_{ho} + \boldsymbol{t_j W_T})$$

where t_j is the topic embedding vector and W_T the corresponding matrix. The authors add the topic to each LSTM gate. Thus the LSTM gates can learn if the topic information in the working memory is still relevant. In AM datasets, words in sentences or argument components are only assigned to one specific topic, but modifications seem possible here. The CLSTM architecture is used in the work of Stab et al. [170] for AId. It drastically improved the performance compared to the standard LSTM. Further, we used the CLSTM as a baseline in our work (see Chapter 2).

1.2.3.4 LSTM with Attention

The standard LSTM has a bottleneck because the feature extraction step needs to represent the entire input sequence $x_1, x_2, \ldots, x_j$ as a single vector c_j. This can cause information loss for long sequences as all information must be compressed into c_j. The *attention mechanism* helps to look at all hidden states h_j for making predictions. A "small" neural network is used to learn which hidden states to *attend* to and by how much. The context vector c_j is now a linear combination of the hidden values h_j weighted by the (learned) attention values a_j

topic	sentence and importance weighting														
school uniforms	forcing	students	to	wear	the	same	clothes	violates	the	students	right	to	freedom	of	expression
	.048	.092	.055	.092	.045	.034	.085	.085	.045	.092	.035	.055	.088	.057	.093
nuclear energy	forcing	students	to	wear	the	same	clothes	violates	the	students	right	to	freedom	of	expression
	.053	.058	.078	.045	.104	.050	.047	.068	.104	.058	.087	.078	.047	.074	.047

Figure 1.4: The attention LSTM weights based on the input words of a sentence and a topic. The first row indicates high relevance on some topic specific words, whereas the second row has a lower relevance on these words [169].

$$c_j = \sum_{t=1}^{T} a_{tj} h_j \tag{1.12}$$

The model [4] was first used for neural machine translation but then spread widely for other tasks. The LSTM with attention architecture has also found its way into AId [170]. Here the attention-based CLSTM learns an importance weighting of the input words of an argument depending on the given topic (see Fig. 1.4).

As the dataset sizes in Natural Language Processing (NLP) steadily increase, the two remaining limitations in RNN's - the *inductive bias* and the *sequential processing* become more of a problem. The sequence models follow the Markov property that each state is assumed to depend only on the previously seen state. This "hardwired bias" of temporal invariance makes the RNN data efficient to train. However, with more data, the inductive bias is increasingly becoming a restriction [144]. The sequential processing makes it impossible to parallelize the architecture since the state s_j depends on the previously computed hidden state s_{j-1}. The Transformer [181] architecture in Section 1.2.4 allows parallelization and further eliminates the recency bias from the RNN architecture.

1.2.4 The Transformer

The Transformer [181] is a self-attention-based architecture, that combined with Bidirectional Encoder Representations from Transformers (BERT) (see Section 1.3), started a new era in NLP and also in most of the related subfields including AM. It allows the processing of whole sequences rather than word by word. This enables parallel processing of all tokens in a sequence and avoids the long dependency issues from RNNs (see Section 1.2.3). The *self-attention* mechanism computes weighting scores inside the sequence and provides information about the relationship between different words and tokens. *Positional embeddings* provide the positional information from RNNs, which encode the location of a word or a token inside a sentence and replace the recurrence mechanism of RNNs. The architecture consists of an encoder and decoder part. The split architecture is necessary for sequence-to-sequence tasks such as machine translation. Here we focus on the encoder part used as a *feature extractor* (see Fig. 1.5). The encoder of the Transformer

architecture consists of three major parts: *Input Embedding and Positional Encoding*, *Multi-Head-Attention* and a *Fully Connected Layer* (Feed Foward) (see Section 1.2.2.1).

1.2.4.1 Input Embedding and Positional Encoding

As an input embedding, pretrained embeddings such as Word2Vec [110] (see Section 1.2.2.2) are an option. However, the original implementation trained embeddings "on the fly", which ensures that out-of-vocabulary words do not occur. As positional encodings learned-[53] and fix encodings (such as sinusoids) are possible. The input embedding and the positional encoding are combined by summarizing each dimension.

1.2.4.2 Multi-Head Self Attention

The self-attention, which is the central contribution of the Transformer architecture, is defined as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_m}})V \tag{1.13}$$

$$Q = XW_Q \quad K = XW_K \quad V = XW_V$$

$$X \in \mathbb{R}^{n \times d_{in}}, W\_ \in \mathbb{R}^{d_{in} \times d_m}, Q, K, V \in \mathbb{R}^{n \times d_m}, d_m \in \mathbb{R}$$

where X is the combined input embedding of a sequence of length n, $W\_$ are the weight matrices. Q, K, V are the queries, keys and values of the input and weight matrices' corresponding matrix product, d_m is a predefined dimension of the model, and d_{in} is the input embedding dimension. The attention score then indicates how the inputs interact with each other. In the Transformer architecture, multiple attention heads can learn diverse relationships between the inputs.

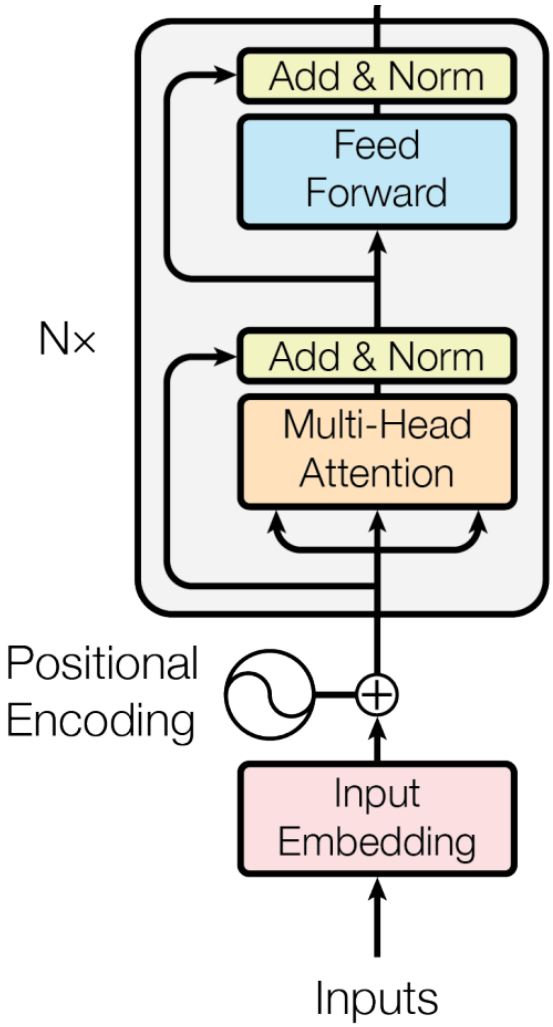

Figure 1.5: The architecture of the Transformer Encoder [181].

1.3 Transfer Learning in NLP

The "ImageNet moment" in Computer Vision (CV) in 2012 describes the moment that a deep neural network (AlexNet) [87] designed by Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton performed 41% better on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [153] than the next competitor, which triggered the deep learning explosion in Machine Learning (ML) research. This moment also allowed a breakthrough of equal importance in Transfer Learning; the weights learned in deep neural networks for ImageNet could be used to initialize neural networks for completely other CV datasets and improve the performance significantly compared to standard initializations such as the "Xavier" weight initialization [60] or the "He" weight initialization [72].

Transfer Learning is a method of extracting knowledge (e.g., representations) from a *source setting* and applying it to another *target setting*. Different Transfer Learning methods have driven major improvements across ML throughout history. These methods can be further classified into three variations:

- **Domain Adaptation:** Source and target task are similar, but labeled data is only available in the source domain (e.g., a particular text genre).
- **Multi-task learning:** Source and target task are different; training occurs parallel on both tasks.
- **Sequential Transfer Learning:** Source and target task are different; *pretraining* occurs first on the source task, and afterward, the model is further *finetuned* on the target task.

We developed architectures and designed pipelines, based on *sequential Transfer Learning* in Chapters 2-7. The source tasks are masked-language modeling (MLM) and next sentence prediction (NSP) (see Section 1.3.2). The target task is either *Argument Identification (AId)* (Chapters 2, 3), *Argument Retrieval (AR)* (Chapter 5), or *Argument Quality (AQ)* (Chapters 6, 7). In Chapters 3 and 7, we develop a technique for AId and Emotion Detection, based on *domain adaptation*. In AId the source domain is heterogeneous text from blogs, newspaper articles, or debate portals, and the target domain is scientific reviews from computer science conferences. For Emotion Detection, we use a combination of seven emotion corpora as the source domain and arguments as the target domain.

In the following subsections, we focus on introducing *sequential Transfer Learning* as these methods are of major relevance for the remaining chapters of the thesis. The pretraining occurs on large source corpora and the finetuning happens on smaller target domain datasets.

Word vectors [110, 128] (see Section Section 1.2.2.2) can be interpreted as sequential Transfer Learning in Natural Language Processing (NLP). They are trained on a large (unsupervised) corpus and are then used to initialize the first layer of neural networks. Word vectors significantly boost prediction accuracy for target tasks with limited target data, such as *AId* or *AQ* (see Section 1.4.2 and 1.4.4). However, they only incorporate

knowledge in the first layer of the model; other network layers still require a relatively large corpus and extensive training for a good performance on the target task.

Language has complex phenomena such as agreement, negation, compositionality, anaphora, polysemy, long-term dependencies and much more; therefore, a more expressive way of Transfer Learning is required. Transfer Learning was of limited use in NLP for recent years (compared to other disciplines, e.g., computer vision) as no Transfer Learning technique could capture this higher-level phenomena.

Recent techniques such as ULMFiT [78], ELMo [130], OpenAI's GPT [139], and Bidirectional Encoder Representations from Transformers (BERT) [33] all have one fundamental paradigm change: instead of initializing only the first layers of the models, they are pretraining the entire model with hierarchical representations on an unsupervised task. Section 1.3.1 discusses different NLP tasks that offer large corpora for the pretraining of deep neural networks. Section 1.3.2, will introduce the BERT [33] architecture which achieves state-of-the-art performances on most Argument Mining (AM) tasks [49, 50, 64, 179].

1.3.1 Pretraining

NLP researchers recently focused on deriving a suitable task for sequential Transfer Learning. As possible contenders there are multiple tasks:

- **Reading comprehension** is the task of answering questions about a paragraph. The most important corpora currently are the Stanford Question Answering Dataset (SQuAD) [141]. This dataset contains more than 100,000 question-answer pairs and require the model to predict a span of the paragraph as an answer.
- **Natural language inference** is the task of identifying the relation (neutral, entailment, and contradiction) between a piece of text and a hypothesis. The dataset for this task is the Stanford Natural Language Inference (SNLI) corpus [17].
- **Machine translation** translates text from one language into another language. It is the most studied task in NLP and accumulated vast datasets over time. CCMatrix [164] contains 4.5 billion parallel sentences in 576 language pairs.
- **Language modeling (LM)** predicts the next word given its previous words in a sentence. As the task is entirely unsupervised, basically any number of sentences can be used for training. The "Colossal Clean Crawled Corpus" (C4) [140] contains over 180 billion target tokens.

Recent research indicates that language modeling (LM) captures many facets of language relevant for downstream tasks, such as hierarchical relations [66], sentiment [138], and long-term dependencies [98]. It was also shown that LM requires less training data for syntactic tasks than other tasks such as translation, skip-thoughts, and autoencoding [193]. The most significant benefit of LM is that training data is free with any text corpus. In the future, the amount of available text on the internet will further increase drastically.

1.3.2 BERT

Bidirectional Encoder Representations from Transformers (BERT) is based on stacking multiple encoders of the original Transformer architecture (see Section 1.2.4) [181]. In order to make BERT flexible to a large variety of target tasks, the model can handle both a single sentence and a pair of sentences. In most of our work [49, 48, 50, 47, 83] we build upon this flexibility by injecting the argument and the topic as an input. The language model is pretrained on the BooksCorpus (800 million words) [194] and English Wikipedia (2,500 million words). BERT does not use the traditional language modeling (LM) task (see the previous subsection). Instead the authors use two unsupervised tasks described in the following.

Masked-language modeling (MLM) is the task of masking some percentage of input sentence tokens at random and then predicting those masked tokens. This allows a bidirectional model that can "see in both directions" compared to standard LM, where the prediction is left-to-right or right-to-left. The aggregation of both LMs in a multi-layered context would allow the model to predict the target word trivially.

Next Sentence Prediction (NSP) is the second task required for the model to understand *sentence relationships*. For each pretraining example, 50% of the time, the following actual sentence in the text was chosen (labeled as *IsNext*); 50% of the cases, a random sentence from the corpus was chosen (labeled as *NotNext*). The model then is trained to classify the sentence pairs correctly. The authors showed that next sentence prediction (NSP) is beneficial for a good performance on the SQuAD [141] and Stanford Natural Language Inference (SNLI) [17] corpora. As we report in Chapter 2, the learned sentence relationship can also be used for Argument Identification (AId) in the context of a topic.

BERT achieved with a single architecture, state of the art results on 11 diverse Natural Language Processing (NLP) tasks in 2018. A wide range of other tasks followed and currently Transfer Learning in the form of pretrained models has become ubiquitous in NLP.

1.4 Argument Mining

Until now, we discussed classical feature extraction, word embeddings such as Word2Vec, and representation learning architectures based on deep neural networks (see Section 1.2). Combined with Transfer Learning (see Section 1.3), these architectures and representations are the foundations of Argument Mining (AM) based on Machine Learning (ML) techniques.

In this Section we present and discuss the theoretical background of AM, show possible application areas, and introduce the AM tasks in the main chapters of the thesis in detail. First, we briefly present the role of argumentation in the society and define argumentation generally. Furthermore, we present how argumentation schemes can define and represent argument components within the argumentation process. Afterward, we discuss the Argument Identification (AId) task, the central task in Chapters 2, 3 and 4. In the third subsection, we discuss the Argument Retrieval (AR) task, the main objective in Chapter 5. In the last subsection, we introduce the Argument Quality (AQ) task, the problem of interest in Chapters 6 and 7.

1.4.1 Theoretical Background

Lately, humanity scholars critique the assumption that reasoning generates better decision-making. By discovering widespread, systematic, and predictable cognitive biases and dysfunctions such as the well-known confirmation bias (new information that confirms our beliefs are enhanced, whereas contrary information is dismissed), researchers think about why we evolved as reasoning creatures. Mercier and Sperber [107] come to the following conclusion in their work:

"Reasoning contributes to the effectiveness and reliability of communication by enabling communicators to argue for their claim and by enabling addressees to assess these arguments. It thus increases both in quantity and in epistemic quality the information humans are able to share."

They claim that reasoning evolved to grant humans better ways of arguing and subsequently better ways of communication. For them, the confirmation bias is a trait that helps people craft good arguments (by filtering out counterproductive information). So rather than being a deficiency when evaluated on decision making, the confirmation bias is helpful to promote better argumentation, communication, and therefore survival.

As we motivated **argumentation** as a precondition of reliable communication, we present a range of definitions of the term. For MacEwan (1898) [104], *argumentation is the process of proving or disproving a proposition. Its purpose is to induce a new belief, to establish truth or combat error in the mind of another.* Ketcham (1925) [84] defines argumentation as *the art of persuading others to think or act in a definite way. It includes all writing and speaking which is persuasive in form.* There exist a plethora of definitions; most define the purpose of argumentation as the *persuasion of others.*

O'Keefe [122] defines persuasion as *a successful intentional effort at influencing another's mental state through communication in a circumstance in which the persuadee has some*

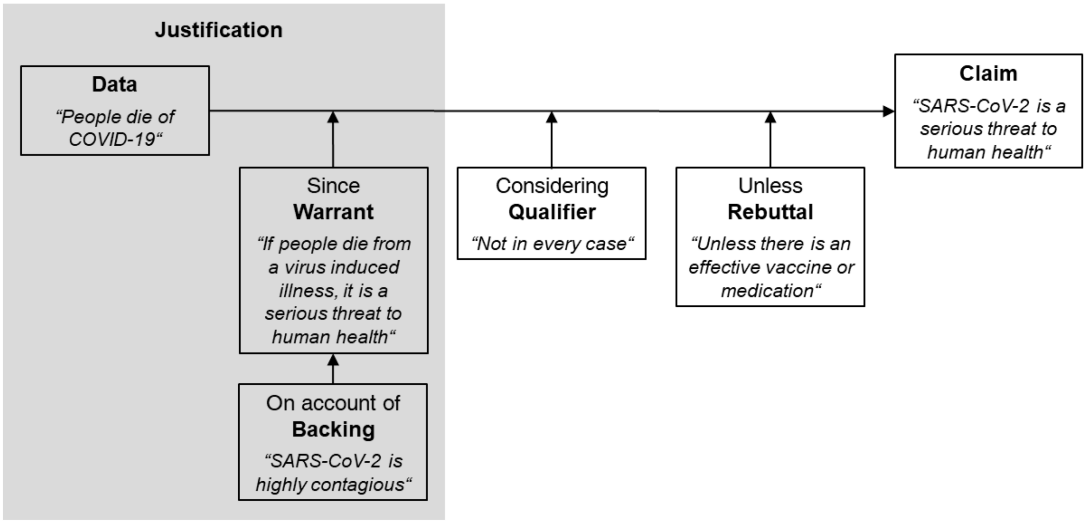

Figure 1.6: An Example of the Toulmin argumentatiom scheme [9].

measure of freedom. Communication as a means of persuasion considers non-reasoned techniques of influence, such as pathos, an emotional appeal to the opponent [13, 143], or ethos, which appeals to the presenter's credibility [143].

1.4.1.1 Argumentation Scheme

Scholars created various concepts of arguments definitions in the following, also called argumentation schemes or argument schemes. The typical and most spread definition is that an *argument is a claim supported by reasons* [162]. Toulmin developed a more complex argument scheme that captures fine-grained roles [177]:

- **Claim** is an assertion put public for general acceptance
- **Data** stands for the evidence to establish the *claim*
- **Warrant** takes the role to justify the logical inference from the *data* to the *claim*
- **Backing** assures the legitimacy of the *warrant*
- **Qualifier** specifies the degree of certainty under which the claim should be accepted
- **Rebuttal** presents settings in which the *claim* might be overcome

In ordinary argumentative discourses, the fine-grained roles of the Toulmin argumentation scheme are often implicit. E.g., the *Warrant* in Fig. 1.6 is known to (nearly) everyone. Humans have a broad background- and contextual knowledge, and thus most aspects of such argumentation schemes are omitted in practice. Hence, argumentation schemes are often inapplicable for texts and require further adaption [67]. Even for single

argumentation components such as the *claim*, an empirical analysis across six different datasets showed that the concept is conceptualized quite differently [31] across the corpora. They summarize their work by stating "that the essence of a claim is not much more than a few lexical clues".

In recent work on computational argumentation, argumentation schemes became simpler and more flexible [170, 178]. Often authors only distinguish between persuasive and non-persuasive text units. This simplification enables broader applicability on different text genres, and requires more background knowledge, making AM an exciting task for Transfer Learning (see Section 1.3).

Another central problem with the conceptualization of arguments and argumentation is that there is no agreement on a single argumentation scheme even among argumentation theorists. Van Eemeren et al. [180] states that:

> "as yet, there is no unitary theory of argumentation that encompasses the logical, dialectical, and rhetorical dimensions of argumentation and is universally accepted. The current state of the art in argumentation theory is characterized by the coexistence of a variety of theoretical perspectives and approaches, which differ considerably from each other in conceptualization, scope, and theoretical refinement."

This concurrency caused an accumulation of rather small datasets (compared to other Natural Language Processing (NLP) areas such as Machine Translation) with different argumentation schemes.

1.4.2 Argument Identification

Argument Identification (AId) is the task of assigning textual units (e.g., words, sentences or text sections) their argumentative structure (see Section 1.4.1.1 on argumentation schemes). The task can be either achieved by a manual argument analysis from a human analyst or by automatic argument extraction methods. A wide range of supporting tools for the manual analysis have been developed, such as Araucaria [150], Rationale [54], OVA [12], and Carneades [61]. These tools are helpful for analyzing a few texts, or for annotation. Large-scale data, or certainly real-time data, require an automatic analysis. Since the AId task has generated a great deal of research and further sub-tasks, work in the area can be assigned in one or more categories:

1. **Isolated AId**: The goal is to separate persuasive text parts from non-persuasive parts

2. **Intrinsic AId**: In this setting, we assign sentences intrinsic attributes such as "is sentence X a premise?" or "is sentence X a claim?"

3. **Relational AId**: This task involves explicit relationships, either between multiple sentences such as "is sentence X an argument for topic T? or between different fine-grained argument roles in an argumentation schema such as "is X a premise of claim Y?

The aforementioned division increases in complexity, starting from isolated AId, where additional context is often unnecessary. Whereas relational AId requires either background information or a precise capture of the text. The different sub-tasks are often performed sequentially, e.g., first, a sentence is classified as an argument, and afterward, an algorithm decides the fine-grained role [116]. This inter-dependency between the tasks motivated multi-objective learning approaches, where all tasks are learned and performed simultaneously [118, 163, 77, 52].

Isolated AId can not give us a detailed picture of the argumentative structure of a text. Nonetheless, it can be helpful for the classification of whole documents based on the *amount* of argumentative content. Moens et al. [116] designed one of the first approaches to AId. Their work uses the Araucaria corpus [146], which contains newspapers, parliamentary records, court reports, magazines, and online discussion boards. They first split the text into sentences and then used various handcrafted features, such as semantic, syntactic and lexical properties (see Section 1.2.1), to classify each sentence as *argumentative* or *non-argumentative*. They achieved an accuracy of 0.74 using a multinomial naive Bayes classifier trained on word couples, verbs, and other features.

A broad range of articles proposes using a two-step approach for AId. In the fist step, a classifier (e.g. a SVM) is used to decide if a sentence is argumentative or not. In the second step, segments of the argument are classified as either premises or claims [62, 158]. Even though these results are encouraging, the classifications carried out only refer to the intrinsic features of the sentence and stand in no further relationship. They may be part of an argument in one context, but in another context they are not. Carstens and Toni [22] raise the point that for a reasonable argument, a relation must first be specified, and in a second step it must be checked if the argument applies to the relation. They give the following example in their work:

1. Nigel Farage has attended private school and used to work as a banker in the City.

On his education and professional past, this is rather a fact than argumentative. If we add further context information, the situation changes:

2. Nigel Farage understands the common folks; he is the face of UKIP, the people's army!

The first sentence can be interpreted as an attack on the claim that Nigel Farage understands the ordinary folks. The conclusion that a private school person can not understand the common people, is not stated explicitly here, but one can infer it.

The example showed that the identification of arguments in isolation might lead to results that are not correct if more context or background knowledge is considered.

Saint-Dizier [155, 154] explore various Argument Mining (AM) corpora on the necessity of domain knowledge and show that in about 75% of the cases, contextual knowledge is required for a reliable AId concerning a disputed topic. Opitz and Frank [125] show that classifiers focus more on the *context* of an argument than on the *content* of an argument. This behavior leads to a solid performance when arguments appear near the associated context. Nevertheless, in a cross-document analysis, such systems can be easily fooled.

Due to the dependency of the context, much research in AId occurred in specialized domains. The general domain-divisional application of AM is problematic. In most domains, the context is captured differently, and models trained in one domain can prove not applicable in another domain. Therefore researchers focused their techniques on certain domains such as persuasive essays [168, 185, 121, 129, 119, 25, 24, 167, 21], tweets [159, 38, 16, 81, 160, 80, 117, 161], legal texts [192, 135, 183] or debates and dialogues [99, 71, 23, 39, 173, 149, 156, 6, 23, 173].

Stab et al. [170] showed that crowd workers could apply their argumentation scheme reliable to sentences from arbitrary Web texts. It consists of the three classes: *supporting-argumentative*, *opposing-argumentative*, and *non-argumentative*; all related to a controversial topic. Many domains were included, such as news reports, editorials, blogs, debate forums, and encyclopedia articles. The resulting UKP Sentential AM Corpus of Stab et al. [170] contains 25,492 sentences from eight controversial topics across the different aforementioned domains. As a feature extractor and classifier, they used the Contextual Long Short-Term Memory (CLSTM) architecture [56] and the attention-based Long Short-Term Memory (LSTM) architecture [74] (for more information regarding CLSTM or LSTM, see Section 1.2.3). They achieved a Macro F_1 score of 64% in the two-class setting (argumentative vs. non-argumentative) and a Macro F_1 of 42% in the three-class setting.

Instead of using the context inside a document or sentence, our work [49] (see Chapter 2) focused on incorporating general external context information in the form of pre-trained Natural Language Processing (NLP) models (see Section 1.3.2) or in the form of Knowledge Graph embeddings [97]. Our evaluation is based on the UKP Corpus. We highlight that external context information of the topic and the sentence can drastically improve the classification performance. In the two-class setting, our best model achieves a Macro F_1 score of 81%, and in the three class setting a Macro F_1 score of 69%, which considerably improved the reliability of the classification.

With the success of our approach [49], we adopted it to the domain of scientific peer-reviews. Peer reviews are central in modern research and were so far underrepresented in AM. Our work [48] (see Chapter 3), found that arguments used in the peer-review process differ from arguments in other domains. Therefore the transfer of knowledge is even with external knowledge difficult. Consequently, we provide the AM community with a new peer-review dataset from different computer science conferences that captures the essence of arguments in this domain. In an extensive empirical evaluation, we show that AId can be reliably used on our corpus. We furthermore show that the extracted arguments are decisive for the publication decision.

1.4.3 Argument Retrieval

Argument Retrieval (AR) also known as *Argument Search*, provides users an overview of viewpoints and arguments regarding a particular topic or claim. The task has a high potential for interdisciplinary use. Lawyers are often in the situation that they want to find arguments that guide their rhetoric in a trial, and they want to be prepared for counterarguments of the opposing faction. Politicians must know the public's viewpoints to estimate how specific legislative measurements are perceived in society.

One of the earliest approaches in AR, Levy et al. [101], designed a system specifically for detecting topic-dependent claims from Wikipedia. The MARGOT system[2] [101] was trained on a corpus of 547 Wikipedia articles. It was used and evaluated on datasets from various genres such as persuasive essays and social media, with encouraging performance.

Args.me[3] [188] is based on debates found on five of the largest debate portals. Their system relies on the pre-structured arguments from these sources and is not generally applicable to web texts. They use Apache UIMA and Lucene for indexing, querying, retrieving, ranking, and presenting the documents and extracted arguments. A manual evaluation study discovered that their systems could find 71% of the expert arguments among the top 50 ranked arguments. However, 47% of the received sentences were either not an argument, was nonsensical, or had the wrong position, meaning that while the system's coverage is high, precision is still a problem.

Summetix (formerly known as ArgumenText)[4] [166, 32] is an Argument Retrieval System (ARS) that is built upon the English part of the CommonCrawl[5] Web corpus. It contains an extensive collection of arbitrary Web texts. The ARS consists of two parts, an offline processing phase, where the documents from the corpus are segmented into sentences and indexed with specific topics, and an online processing phase where Argument Identification (AId) [170] is performed on the selected sentences. In a manual evaluation of the system, the top-ranked results are compared with arguments curated on an online debate portal. The system has achieved high coverage of 89% with regarding the curated lists. Like Args.me [188], the precision is an issue, with slightly less than half of the arguments being irrelevant or misclassified in their position to the topic. Furthermore, the top-ranked results often contain similar semantic arguments, as their method does not filter out these.

Dumani et al. [36, 35] pointes out that semantically similar premises are often formulated differently. An ARS should avoid these duplicates and therefore requires some form of similarity measure and clustering. Instead of relying on an AId component in the ARS, they propose to decouple the task of AR and AId. In their ARS, users can formulate queries to access relevant *premises* for a given *claim*. An example of a claim related to *energy* could be *"We should abandon nuclear energy"* and a relevant supporting premise, e.g., *"Accidents caused by atom reactors have longstanding negative impacts"*. As noted by Dumani et al. [36, 35], a sole similarity-based approach can not automatically be

[2]MARGOT: Mining Arguments from Text. `http://margot.disi.unibo.it/`
[3]Args.me: `https://www.args.me/index.html`
[4]Summetix: `https://www.argumentext.de/`
[5]CommonCrawl Web corpus: `http://commoncrawl.org/`

associated with the relevance of a premise. The authors recommend using similarity between a query claim and a result claim, with the latter being associated with multiple premises. The assignments between premises and result claims are extracted from multiple debate portals. The evaluation is performed on a subset consisting of 1195 triples (query claim, result claim, result premise), where premises were annotated as "very relevant", "relevant" and "not relevant" regarding the query claim. Additionally, the 528 triples categorized as "relevant" or "very relevant" were clustered by annotators. They used a clustering extended, α-nDCG (normalized Discounted Cumulative Gain) [27] scores, as an evaluation measure. Their best ARS achieved a mean nDCG@5 of 45.5% and a mean nDCG@10 of 48.7%, which significantly improved upon the BM25F baseline. Recently they also added a quality-aware ranking step in their ARS [37].

In our work [50] (see Chapter 5), we build on the ARS from Dumani et al. [36, 35] by decreasing the required ground truth information between the result claims and the premises. This lowers the argument corpora requirements and therefore makes the ARS broadly applicable. Instead of relying on the similarities between the result claims and the premises, we advocate using a pretrained language model (see Section 1.3.2) to learn the relevance between premises and query claims. Thus we remove one step of the previous AR and can evaluate the relevance directly on the premises and the query claims. We use a negative-sampling method based on the premise similarity for generating *negative samples*, and the premise-claim pairs of Dumani et al. [35] for the *positives*. In this setting, the relevance filter can learn the fine-grained semantic differences between relevant and non-relevant premises regarding the claim.

We also propose a novel diversity component, which selects a representative subset of diverse premises from the relevant ones. Our best ARS achieves an nDCG@5 at 47.5% and a mean nDCG@10 of 52.6%, which significantly increases the ranking quality of premises compared to previous work.

Touché 2020 [14] was organized as a collaborative platform for research in AR. The platform introduced a competition and datasets for two tasks: (1) supporting individuals in finding arguments on socially important topics and (2) supporting individuals with arguments on everyday personal decisions. In 2021 the collaborative platform [189] organized another competition. The first task focuses on AR for controversial questions. The second task aims at supporting users facing a choice problem. Given a comparative question, the task is to retrieve and rank documents to help answer these questions. A more specialized AR task is retrieving the best counterargument for a given argument. Wachsmuth et al. [190] created an corpus with over 6000 argument-counterargument pairs taken from 1069 debates. They created eight retrieval tasks with different complexity based on the new dataset. Multiple similarity measures, such as the Manhattan and Jaccard similarity, are evaluated on word embeddings and handcrafted features.

Name	Sentences	Topics	Domain	Quality Notion	Abs.
Wachsmuth [187]	320	16	Debate Portal	15 Dimensions	Yes
UKPRank [68]	1,000	32	Debate Portal	Convincingness	Yes
SwanRank [68]	5,300	4	Debate Portal	Interpretability	Yes
IBM-ArgQ [175]	6,300	11	Crowd Coll.	Recommended	Yes
IBM-EviConv [58]	8,000	118	Crowd Coll.	Evidence	No
IBM-ArgQ-Pairs [175]	14,000	11	Crowd Coll.	Recommended	No
UKPArgAll [68]	16,000	32	Debate Portal	Convincingness	No
Gretz [64]	30,000	71	Crowd Coll.	Recommended	Yes
Potash [133]	71,840	3,439	Web documents	Convincingness	No

Table 1.3: Overview of different Argument Quality Datsets

1.4.4 Argument Quality

Argument Quality (AQ) (sometimes also called argument strength) is a further sub-task in Argument Mining (AM). AQ is often captured differently due to its high subjectivity. It can be measured as a continuous score (absolutely) or in relation to other arguments. Wachsmuth et al. [187] provide a corpus with 320 arguments, annotated for 15 fine-grained argument dimensions taken from theory. They categorize the quality dimensions into three main quality aspects:

- *Logical quality* in terms of the cogency or strength of an argument
- *Rhetorical quality* in terms of the persuasive effect of an argument or argumentation
- *Dialectical quality* in terms of the reasonableness of argumentation for resolving issues

The authors clarify in their work that practical approaches can help focus on the simplification of theory and that AQ theory can guide practice. Based on their corpus from 2017, Wachsmuth et al. [191] trained a linear SVM based on eight handcrafted feature types to predict the 15 fine-grained argument dimensions. They found that the modeling of logical and dialectical dimensions, in terms of subjectiveness, sentiment, pronoun usage, and similar are possible on arguments included in the corpus. Due to the limited corpus size, it was hard to find complex features that robustly predicted the overall AQ.

Research in AQ recently created larger corpora that focused on a single "overall" quality score (see Table 1.3).

Swanson et al. [173] developed an automatic regression method to estimate point-wise AQ. They constructed the dataset *SwanRank* with over 5k arguments labeled in the range of [0, 1], where a 1 indicates that an argument can be easily interpreted in a given dialogue. They used linear regression, ordinary Kriging, and Support-vector machines (SVM) as regression algorithms on "handcrafted" features such as the sentence length or discourse and dialogue features. They evaluated the models on evaluation measures such

as the Root Mean Squared Error. All features paired with an SVM performed best and improved upon the lexical n-grams baseline for all topics.

Other approaches constructed argument corpora based on the relative- or absolute *convincingness* [69, 70, 133, 132]. *UKPConvArgRank* (absolute) and *UKPConvArgAll* (relative) contain 1k labeled arguments, and 16k labeled argument-pairs. Habernal et al. [70] developed two architectures for the absolute and relative AQ tasks on the UKP datasets. The first model is an SVM with an radial basis function (RBF) kernel based on a large set of rich linguistic features. The second model is a bidirectional Long Short-Term Memory (BiLSTM) architecture with Global Vectors (GloVe) embeddings [128] (see Section 1.2). The "classical" SVM outperforms the BiLSTM in both tasks, with a Macro F_1 score of 78% (SVM) and 76% (BiLSTM) on the relative task and a Pearson correlation of 35.1% (SVM) and 27.0% (BiLSTM) on the absolute task.

Potash [133] uses a sum-of-word-embedding approach based on GloVe word embeddings [128]. The word embeddings serve as an input to a Multilayer Perceptron (MLP) with three layers of sequentially decreasing size. They evaluate their model on their dataset and the UKP dataset. This architecture could further increase the performance of the absolute AQ task on the UKP corpus to a Pearson value of 48%.

Gleize et al. [58] dataset *IBM-EviConv* focus on ranking *evidence's* convincingness. They used a siamese network based on a BiLSTM with attention and trainable Word2Vec embeddings. They evaluated their architecture on the UKP datasets and their own and could achieve in both scenarios new state-of-the-art results. Gretz et al. [64] and Toledo et al. [175] created their corpora of 30k and 6.3k arguments by asking annotators if they would recommend a friend to use the argument in a speech supporting/contesting the topic, regardless of their personal opinion. Both finetuned Bidirectional Encoder Representations from Transformers (BERT) [33] for the absolute AQ estimation regression task and could achieve new state-of-the-art results. Toledo et al. [175] concatenated the last four layers of the model output to obtain the embedding vector for the regression layer. In contrast Gretz et al. [64] use the last layer directly as the vector for the regression unit.

Contrary to the discussed work that primarly focuses on isolated datasets and neglects the interactions with related AM tasks, our work [47] (see Chapter 7) estimates the AQ models applicability in more challenging scenarios. We advocate for *cross dataset* evaluation without additional finetuning on the other corpora. Also, we approach AQ from two other angles: We assess the interplay with related AM tasks and the impact of emotions on the perceived argument strength.

Our other work [83] (see Chapter 6) in this area evaluates uncertainty-based active learning methods on the task of relative AQ. We evaluated our approaches on the corpora of Toledo et al. [175] and the UKPConvArgAll [68] corpora. We use different data selection strategies and benchmark them against the random data selection, which serves as a baseline. We measure the model uncertainty by approximating the Bayesian inference through dropout in neural networks [51]. In our work, we point out issues in AQ that prevent the efficient use of uncertainty-based active learning methods and give insights on how they can be addressed in future work.

1.4.5 Related Areas & Applications

Two areas related to Argument Mining (AM) are the areas of *opinion mining* and *sentiment analysis*. Opinion mining is defined as "the computational study of opinions, sentiments, and emotions expressed in text" [102]. Sentiment analysis is limited to positive and negative views in a text span, whereas opinion mining may enclose a great variety of opinions. Connections between opinion, sentiment, and arguments are further explained in the work of Hogenboom et al. [75]. They point out that expressions of sentiment in text can be used as an indicator for argumentative structures. The sentiment analysis and opinion mining area discuss *what* opinions are presented in a text. In comparison, AM focuses on *why* people have these sentiments or opinions. Another related area is *controversy detection*. In controversy detection, the focus is on recognizing controversial topics and text spans where opposing standpoints are discussed. Similar to AM, controversy detection primarily targets specific text domains such as Wikipedia articles [85] or news articles [26, 3]. The degree of controversy for the news articles is calculated by the volume of positive and negative sentiment and the difference between them. The two remaining areas of *citation mining* and *argumentative zoning* are closely related to Chapter 3, where we deal with peer-reviews. Citation mining involves labeling citations in scientific papers and reviews based on their rhetorical role. Similar to arguments in AM, citations often have either a supporting role, e.g., citing assisting related work, or an opposing role, such as highlighting a gap or a limitation of related work. In Piao et al. [131], the authors use existing semantic lexical resources and Natural Language Processing (NLP) tools to identify the author's opinions towards the work they cite. They classify the citations attitudes such as positive/negative or approval/disapproval. *Argumentative zoning* is the classification of sentences by their argumentative and rhetorical part in a scientific paper. Possible categories are the comparison of methods or results and criticism or support for previous work. Teufel et al. [174] developed an annotation scheme of 14 mutually exclusive classes to classify sentences in papers. Another work in the area is from Merity et al. [108], where they used a maximum entropy classifier to categorize sentences into seven rhetorical structures. Their method was evaluated on 48 computational linguistics papers taken from proceedings of multiple conferences.

1.4.5.1 Applications of Computational Argumentation

Nowadays, argumentation and communication have become more challenging since disputes are often global, involve multiple stakeholders, and require reasoning from specialists that are not easy to follow for everyone. A first step to ease this situation is mining and providing arguments from diverse text areas such as social media, newspapers, debate portals, and scientific literature. AM is not limited to written monologs or dialogs. It can also be used on spoken monologs and dialogs such as political debates or panel discussions.

Project Debater [165] goes a step further; it is an autonomous debating system that engages in competitive debates with human expert debaters. In June 2018 it debated against Haris Natarajan, one of the world's most decorated debaters, on preschool subsidies. The goal of the project is to provide humans with an AI capable of helping us make

better, more informed decisions.

In general, there are many applications for AM such as Argument Retrieval (AR) [188, 50, 57, 94, 19, 15, 36, 35], intelligent personal assistants [165, 182], fact-checking [2], automated decision making [172, 105, 106], argument summarization [44, 115, 7, 8] and writing support [123, 119, 186]. The underlying task for all these applications is the task of Argument Identification (AId) (see Section 1.4.2), which highlights its importance in the AM research.

1.5 Overview of Contributions

This section, provides an overview of the thesis and locates the publications within the Argument Mining (AM) research area.

- In our work in Chapter 2, we study the task of *Argument Identification (AId)* in heterogeneous text domains and propose Machine Learning (ML) methods, which enables them to capture external context and topic information. Earlier approaches often isolated the tasks of AId from context and topic information or only relied on in-corpus context. This is a problem in the AId setting since the *context* is often more important than the *content* of a text span, especially in a cross-document setting. Our work shows that topic information is crucial for AId since the topic defines the semantic context of an argument. Our evaluation on the data set of Stab et al. [170] highlights that external context information from pre-trained Natural Language Processing (NLP) models and Knowledge Graph embeddings provide a drastic classification performance boost on the AId task compared to previous state-of-the-art approaches.

- In Chapter 3, we propose the application of AId to the domain of *peer-reviewing*. Peer-reviewing is central in modern research and essential for ensuring a high-quality standard of published work. We empirically validate our conjecture that arguments drive the reviewing process in science. Furthermore, we show that domain adaptation from other heterogeneous text domains in AM is only possible to a limited extent as arguments in peer-reviews have their peculiarities. Therefore, we extend the public AM corpora by creating a new peer-review dataset from multiple computer-science conferences. An extensive evaluation shows that fine-tuned AId models can nearly reach human performance on different AId tasks. Additionally, we demonstrate that the extracted arguments play a decisive role in the paper-acceptance decision.

- Chapter 4, describes our research on *relational AId* in the project *ReMLAV* inside the DFG Priority Program *RATIO*. We formalize fine-grained [179] AId as a sequence labeling approach and compare it to previous coarse-grained scenarios [49]. Further, we introduce a novel method for the same-side classification (SSSC) challenge [171].

- Chapter 5 addresses the *Argument Retrieval (AR)* task. After the *AId* step, the end-user wants to retrieve relevant arguments from an argument collection. Our work

focuses on retrieving relevant arguments for a user-defined query claim. Contrary to other work, our approach does not rely on explicit mappings between claims and premises and thus can be applied in an inductive setting, where new premises can be used without manual association of relevant claims. We introduce a new multi-step approach that captures semantic relationships between argument components. In the first step, our pipeline uses a newly designed ML based relevance filter that assigns each premise a relevance value that indicates the suitability, given the query claim. In the second step, we propose a novel diversity component, which selects a representative subset of diverse premises from the relevant ones. Our evaluation shows that our Argument Retrieval System (ARS) significantly improves the ranking quality compared to competitors, even though it does require fewer annotated input data.

- In Chapter 6, we address the *Argument Quality (AQ)* task. We show how uncertainty-based Active Learning (AL) methods can be applied to AQ data sets. Our research highlights difficulties with label-efficient learning on AQ datasets with uncertainty in the annotations. It shows that further requirements are needed for AQ data sets to benefit from AL.

- In Chapter 7, we broaden the perspective for the automatic estimation of AQ. Our work brings together various aspects: First, we evaluate whether AQ models generalize across datasets and domains. That is an essential attribute in practical applications, such as AR. Next, we investigate if other AM tasks are helpful for the target task of AQ estimation. We evaluate the zero-shot performance of *AId* and *Evidence Detection (ED)* regarding AQ. Lastly, we introduce the first dataset that goes beyond the mode of logos. Our corpus additionally captures pathos in the form of emotions in the argument domain. Furthermore, we show that these emotions can be reliably detected.

In conclusion, we think that this work significantly extends the current ML based approaches in AM. In our research articles, we developed approaches for *AId*, *AR*, and *AQ*. We show how AId can be solved with high reliability by combining external context information for both the argument and the topic. In AR, we improve the diversity of retrieved relevant arguments, while relying on less input data than competing approaches. For AQ estimation, we extend the scope of previous work and show generalization capabilities, the interplay with related AM tasks, and the impact of emotions on the perceived argument strength. Furthermore, we investigate on the sample efficiency of modern architectures in AQ and suggest additional requirements for an active-learning-based training pipeline in AQ corpora. Overall, these achievements improve the extraction, discover-ability, and quality estimation of arguments in texts.

2 TACAM: Topic And Context Aware Argument Mining

The chapter includes the following publication:

> **Michael Fromm**, Evgeniy Faerman, and Thomas Seidl. „TACAM: Topic And Context Aware Argument Mining.“ In: *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE. 2019, pp. 99–106

Declaration of Authorship Michael Fromm proposed the research idea, developed and conceptualized it with Evgeniy Faerman, and discussed it with Thomas Seidl. Michael Fromm did the implementation, the design of the architecture and framework. Michael Fromm designed and conducted the experiments and analyzed their results. Michael Fromm and Evgeniy Faerman discussed the results and wrote the manuscript. All authors revised the manuscript.

TACAM: Topic And Context Aware Argument Mining

Michael Fromm
fromm@dbs.ifi.lmu.de
Database Systems and Data Mining, LMU
Munich, Germany

Evgeniy Faerman
faerman@dbs.ifi.lmu.de
Database Systems and Data Mining, LMU
Munich, Germany

Thomas Seidl
seidl@dbs.ifi.lmu.de
Database Systems and Data Mining, LMU
Munich, Germany

ABSTRACT

In this work we address the problem of argument search. The purpose of argument search is the distillation of pro and contra arguments for requested topics from large text corpora. In previous works, the usual approach is to use a standard search engine to extract text parts which are relevant to the given topic and subsequently use an argument recognition algorithm to select arguments from them. The main challenge in the argument recognition task, which is also known as argument mining, is that often sentences containing arguments are structurally similar to purely informative sentences without any stance about the topic. In fact, they only differ semantically. Most approaches use topic or search term information only for the first search step and therefore assume that arguments can be classified independently of a topic. We argue that topic information is crucial for argument mining, since the topic defines the semantic context of an argument. Precisely, we propose different models for the classification of arguments, which take information about a topic of an argument into account. Moreover, to enrich the context of a topic and to let models understand the context of the potential argument better, we integrate information from different external sources such as Knowledge Graphs or pre-trained NLP models. Our evaluation shows that considering topic information, especially in connection with external information, provides a significant performance boost for the argument mining task.

KEYWORDS

transfer learning, argument mining, argument search, natural language processing

ACM Reference Format:
Michael Fromm, Evgeniy Faerman, and Thomas Seidl. 2019. TACAM: Topic And Context Aware Argument Mining. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI '20), October 14–17, 2019, Thessaloniki, Greece.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The main focus of argument search lies on presenting an overview of different standpoints and their justifications to some inquired topic e.g. *cloning* or *minimum wages*. This may be useful in different scenarios, like legal reasoning [44] or decision making processes [35], especially if a topic or a problem is controversial. An automated argument search process could ease much of the manual effort involved in these areas, especially if it can make use of large text databases or even combinations of them. The online argument search in state-of-the-art argument search systems proceeds in two steps [30]:

(1) Some standard text search engine, e.g. [3], extracts relevant text parts from large text corpora using a given topic as a query.
(2) Relevant text parts are analyzed sentence-wise by an argument recognition component which decides for each sentence whether it is an argument and optional about its stance.

Therefore, the core technique in argument search is argument recognition or argument mining [17, 18, 25, 36, 40]. The basis for argument mining is an argument model (to avoid confusion with machine learning models in the following we refer to argument model as argument scheme). An argument scheme formally defines what kind of arguments exist and what their properties and relationships between them are. State-of-the-art argument search systems work with simple argument models without relationships between arguments. Therefore, the common task of a machine learning model is argument recognition or identification.

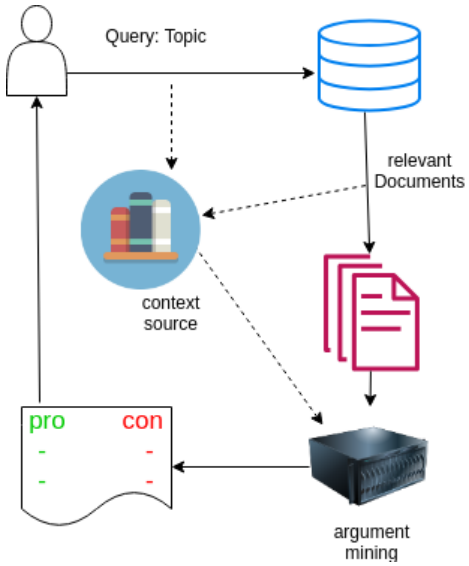

Figure 1: Argument search pipeline with context

WI '20, October 14–17, 2019, Thessaloniki, Greece

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

The classic argument recognition approaches extract arguments from text without taking the topic of the argument into consideration [9, 12, 24]. However, the special characteristic of application of argument recognition in argument search is that there is always a query topic. The query topic carries information about the query context and understanding the context of potential arguments can be crucial for the decision. For instance, if the query is about the usefulness of some medical procedure in the context of medicine, we expect appropriate arguments from the medical doctors and not from people who share their own individual experiences. Thus, if a potential argument follows a particular structure or some special terminology is used, it may increase an argument's chances to be classified as an argument.

Another desirable property of an argument identification approach is to be able to decide **dependent** on a topic, whether a sentence is an argument. To be useful, an argument search engine heavily relies on a large text corpus. The larger the text corpus is, the more probable is the scenario that texts extracted by the search engine contain arguments about different topics, which increases coverage of different topics but further complicates the argument classification. For instance, consider the following example: A user looks for arguments to *Emission Trading System (ETS)* and the following sentence candidates are retrieved by the text search engine:

- ETS sets a clear price on carbon and combats climate change.
- Free trade secures all the advantages of international division of labour.
- UK signals plan to leave EU emissions trading scheme after Brexit.

The first two sentences can be considered as arguments, the third sentence is purely informative and it does not persuade towards any stance. However, if we look more closely at the second sentence, we recognize that this is not an argument about the query topic *Emission Trading*. This is obvious for humans, since we understand that the context of *Free Trade* is different from the *Emission Trading* context, even if both contexts are related to *Trade*. Therefore, the better the machine learning model is able to grasp the context of a topic and of potential arguments at different granularities, the better is the decision the model can make and the more certain it can be about its decisions. Considering relationship between potential argument and topic is different from the classical relation detection task in argument mining. The input to relation detection algorithm are parts of the text, which are already recognized as arguments and presence or absence of some relationships does not affect decision about text parts of being argumentative. In contrast, our approach takes the relationship to the topic into account when deciding whether a text is an argument.

In this work we propose a new approach for argument mining which also takes the topic of potential arguments into account. The overview of our approach is depicted in Figure 1. The standard argument search pipeline looks like the workflow presented in the figure without the context source and the dotted arrows. Our approach enriches the argument candidates with the context and topic information in the classification process. We show how the contextual information about a topic and an argument from different sources like knowledge graphs or pre-trained models can be integrated into our approach. We investigate the benefits of considering the topic and the integration of external knowledge. We summarize our main contributions as follows:

- We present a novel approach for argument classification which takes the topic of the argument into account by extending the methodology introduced by the authors of [34].
- We show how contextual information about topic and argument from different sources like knowledge graphs or pre-trained models can be integrated.
- We demonstrate that considering topics is beneficial for the argument classification, especially in connection with external knowledge.
- We show that our approach is particularly successful if the model has to generalize to unseen topics. Since we cannot expect that available training datasets for argument recognition cover all possible topics, the generalization to unseen topics is an important requirement.
- We present thorough experimental evaluations of our models and comparisons to state-of-the-art methods on a real-world dataset and introduce an additional experimental setting. In this setting we evaluate the ability of different models to classify in the context of topics.

2 RELATED WORK

In general, the main focus in argument mining lies in the recognition of argument components [12, 15, 21, 24, 33] and the detection of relations between them [22, 33]. However, all these approaches which tackle the problem of argument classification do not take information about the specific topic of a given argument into consideration.

At the same time different argumentation schemes of different complexity were proposed in previous works [10, 32, 37, 43]. Since each argumentation scheme contains different numbers of various argument types, this has an implication on machine learning models designed for argument detection, since they have to learn how to identify them. However, as was shown in [7], these argumentation schemes do not generalize well to different types of texts. Concretely, the authors of this work collected datasets used with different argumentation schemes and combined them in a single dataset. Afterwards, they trained a model, which should detect the argument component of type *claim*, which is central in each argument scheme. However, the machine learning models which perform well for single datasets could not achieve good results on this simple binary classification task. Additionally, it was shown that even human annotators often label differently when annotating the same datasets according to complex argumentation schemes. Therefore, the authors came to the conclusion that certain argument components (backing, warrant) as introduced in [37], and other argumentation schemes are often only stated implicitly in common argumentation documents on the internet. In more recent work, argumentation schemes became simpler and more flexible [34, 42]. This enables broader applicability and topic-dependent argument search across multiple text types.

There are various approaches to consider context in argument mining. Hand-crafted features extracted from source text were used for argument classification [21] and relation detection [22]. More related to our work is a method presented in [34]. The authors introduced a dataset with arguments of different text types and topics for each argument. Additionally, they propose two simple argumentation

schemes. The first scheme is a binary decision, aiming at classifying a sentence as argumentative or non-argumentative. In the second scheme there is a distinction between non argumentative sentences and pro and contra arguments. They also propose a model which takes topics into consideration. We extend their work by proposing new architectures and context sources and compare our approach with their method.

There are few approaches which use transfer learning for the argument mining task. In [34] the proposed model is pre-trained on another dataset for argument mining [13], but this approach does not lead to considerable improvement. Parallel to our work, the authors of [38] also use transfer learning with BERT for a new introduced corpus with tagged sequences. However, their model does not generalize to the new topics by design.

Based on recent developments two argument search engines, i.e., *www.args.me* [42] and *www.argumentsearch.com* [31], where a user is able to search a broad range of documents for certain topics, have been developed.

3 PROBLEM SETTING

We model the recognition of argumentative sentences as a classification task. Given a sentence $s = \{s_0, \ldots, s_n\}$ and a topic $t = \{t_0, \ldots, t_k\}$ with $s_i \in \{0,1\}^V$, $t_i \in \{0,1\}^V$ being one-hot encoded vectors, and V being the size of the vocabulary, we seek to classify s as "contra argument" or "pro argument" if the sentence s includes evidence for supporting or opposing the topic t. If the sentence does not contain evidence, it is classified as a "non-argument".

4 METHOD

In contrast to previous approaches, we aim at incorporating context information into the learning procedure when training our models. This way, the models learn which argument properties are especially meaningful in the context of a particular topic and can put a special emphasis on these information for the subsequent classification task. For instance, *emission trading* is a frequently discussed topic, but we would expect the most meaningful arguments about its usefulness coming from particular academic communities. Consequently, by providing topic information in a meaningful way, we enable models e.g. to learn argument structures and vocabulary which are common in those communities. On the other hand we also expect our models to learn how topics are related to their domain specific arguments. Although a sentence might contain topic-specific words it may still be an argument of a different topic. Considering the topic *emission trading* again, relevant arguments are probably more related to climate change than to the stock market, though *trading* is a frequently used term in the latter area. Thus, it is important to *understand* the context of the topic and the context of the potential arguments. Consequently, we propose various approaches to provide context information about topic and potential argument from various external sources. However, as the proposed models should be able to generalize to arbitrary topics, we provide the context information as an additional input to the models. Therefore, all our models aggregate the representation of the potential argument with the representation of the topic.

4.1 Models

4.1.1 *Recurrent Network.*

The first model we propose is a recurrent model for which we use two instances of a BiLSTM [14] model. Precisely, one is used to encode a topic and the other model aims at encoding the potential argument:

$$
\begin{aligned}
x^s &= \{s_1 W^{we}, \ldots, s_n W^{we}\} \\
h^s &= BiLSTM_a(x^s)
\end{aligned}
$$

$$
\begin{aligned}
x^t &= f_{map}(t) \\
x^t &= \{x_1^t W^{te}, \ldots, x_m^t W^{te}\} \\
h^t &= BiLSTM_t(x^t)
\end{aligned}
$$

$$
\begin{aligned}
h_l &= aggr(h^s, h^t) \\
\hat{y} &= softmax(h_l W_{final} + b_{final})
\end{aligned}
$$

We use *word2vec* [19] embeddings $W^{we} \in \mathbb{R}^{V \times d}$ of the given words in a sentence s as input for the argument BiLSTM instance $BiLSTM_a$. However, it is noteworthy that any other kind of word embeddings can be used, too. Furthermore, function f_{map} maps some given topic description t to a sequence of entities x^t. In general, we allow arbitrary information sources to provide topic context. Therefore, f_{map} depends on the information source. In case of describing the relevant entities of t in terms of relevant words, one could use a sequence of word embeddings to encode the topic information. In this case f_{map} would map the relevant words to the corresponding one-hot encoded vectors which, if multiplied with the word embedding matrix W^{we}, serve as input for the topic BiLSTM instance denoted as $BiLSTM_t$. In case of using knowledge graphs as external source of information for the context, f_{map} first examines whether there is an entity with the same name as the whole topic description. Otherwise it maps each word in the topic description to an corresponding entity in the knowledge graph. If there is no such corresponding entity for a particular word, we employ a nearest neighbor search for this word in the word embedding space and finally use a knowledge graph entity which matches to a semantically similar word. Once we found an entity for each word in the topic description, we use the corresponding sequence of knowledge graph entity representations as input for the topic BiLSTM instance. The function *aggr* is used to aggregate topic and argument representations. We evaluate the following aggregation functions:

- Addition: $aggr(h^s, h^t) = h^s + h^t$
- Hadamard product: $aggr(h^s, h^t) = h^s \odot h^t$
- Concatenation: $aggr(h^s, h^t) = concat(h^s, h^t)$

Finally, we use the aggregated representation h_l as input to a dense layer with softmax activation to obtain the classification result $\hat{y}$.

4.1.2 *Attention model.*

We also use a deep bidirectional transformer encoder [39], the architecture which was used in BERT [8]. Specifically, we concatenate argument and topic description and use a special separator token and segment embeddings to distinguish between topic and potential argument. The output of the first special [CLS] token is used as input to the dense classification layer, which predicts the distribution over the classes.

4.2 Context source

As mentioned previously, our models are able to rely on different external sources that may provide the context information. In this work, we experiment with the following sources:

- Shallow **Word Embeddings** [4, 19, 26] are widely used in NLP applications and encode context information implicitly. In fact, the word embeddings are learned such that the representations of words that frequently appear in similar contexts are similar to each other. We use shallow word embeddings trained by word2vec as input to the recurrent model.
- **Knowledge Graphs** model information about the world explicitly in the form of an heterogeneous graph. The entities in the knowledge graph are represented as nodes, and relationships between them as edges of different types. Information in a knowledge graph is represented as triples consisting of subject, predicate and object, where subject and object are entities and predicate stands for the relationship between them. In contrast to information contained in text data, knowledge graphs are structured, i.e., each entity and relationship have a distinct meaning, and the information about the modelled world are distilled in form of facts. These facts can be extracted from texts, different databases or inserted manually. The trustworthiness of these facts in publicly available knowledge graphs is in general very high [23]. In our work we use the english version of the DBpedia knowledge graph, which has about 400 million facts with more than 3.7 million unique entities [16]. We applied TransE [5] to obtain embeddings for the knowledge graph entities. These embeddings are used as input to a recurrent model (alternatively to the word embeddings).
- Fine-Tuning based **Transfer Learning** approaches [8, 28, 29] adapt whole models, that were pre-trained on some (auxiliary) task, to a new problem. This is different from feature-based approaches which provide pre-trained representations [6, 27] and require task-specific architecture for a new problem. We use the weights of pre-trained BERT (Large and Base) [8] models for initializing our 4.1.2 model and train it for the argument classification task.

5 EVALUATION

5.1 Dataset and Evaluation Tasks

For the evaluation we use the UKP Sentential Argument Mining corpus from [34]. The dataset consists of more than 25000 sentences from multiple text types covering eight different topics. It contains a broad range of genres including news reports, editorials, blogs, debate forums and encyclopedia articles which are all related to at least one topic. The topics have been randomly selected from a list[1] of controversial topics. The authors define an argument as a sentence that can be used to oppose or support a given topic. For all models each sentence is truncated to 60 words according to the experiment setting in [34]. Note that in contrast to [34] we use weighted cross-entropy to account for class imbalance.[2] Following [34] we evaluate our approach by performing the following classification tasks:

- Binary classification: whether a sentence is an argument for the given topic.
- Multiclass classification: whether a sentence is supporting, respectively attacking an argument, or is not an argument at all for the given topic.

As suggested in [34], we evaluate all approaches in two different scenarios. In the *In-Topic* scenario each topic is split into training and test data, which leads to arguments of the same topics in both training and test data. The *Cross-Topic* scenario primarily aims at evaluating the generalization of the models, i.e., answering the question how good the performance of the models is on yet unseen topics. Therefore, seven topics are used for training and the remaining one for test. Let us mention that although *Cross-Topic* is the more complex task, it is more relevant for real-world problems: The reason is that in general we cannot expect all possible topic queries to be present in a dataset that is available for training.

5.2 Models

For all tasks we compare the following approaches:

- BiLSTM is a bidirectional LSTM model [14], which does not use topic information
- BiCLSTM is the contextual biderectional LSTM [11]. Topic information is used as an additional input to the gates of an LSTM cell. We use the version from [34] where the topic information is only used at the $i-$ and $c-$gates since this model showed the most promising result in their work.
- TACAM-WE is our recurrent model described in 4.1.1 which uses word embeddings to define the context of the topic
- TACAM-KG is our recurrent model described in 4.1.1 which uses Knowledge Graphs embeddings from DBPedia to define the context of the topic.
- TACAM-BERT Base / TACAM-BERT Large are our attention based models with topic information described in Section 4.1.2. Both model use pre-initialized weights (cf. Section 4.2). TACAM-BERT Base has 1/3 parameters of TACAM-BERT Large .
- CAM-BERT Base/ CAM-BERT Large are similar to TACAM-BERT Base and TACAM-BERT Large models without topic information. These models enrich only potential argument with the context, but do not have access to the topic. Comparing them with their counterparts with topic information enables the evaluation of topic importance.

In our experimental setting we mostly follow the experimental settings suggested in [34]. We use the same train/validation/test splits. The validation set is used to select the hyperparameters and we report Macro F1 scores on test sets. To avoid effects of bad initialization and local minima we train each model 10 times and select the model which performs best on the validation set.

[1] https://www.questia.com/library/controversial-topics

[2] We assume this is a reason we obtained better results for the comparison methods as stated in the original paper.

5.3 In-topic Results

The results of the in-topic argument classification are listed in Table 1. In this setting we do not expect a large improvement by providing topic information since the models have already been trained with arguments of the same topics as in the training set. The results in Table 1 reflect our expectations: we can slightly improve the classification results for the more complex multiclass classification problem. However, we see a relative increase of about 10% for the two-classes and 20% for the three-classes classification problem by using context information from transfer learning. Therefore, we conclude that contextual information about potential arguments is important and since the topics are diverse, the model is able to learn argument structure for each topic.

	Method	
two-class	BiLSTM	0.74
	BiCLSTM	0.74
	TACAM-WE	0.74
	TACAM-KG	0.73
	CAM-BERT Base	0.79
	TACAM-BERT Base	**0.81**
	CAM-BERT Large	0.80
	TACAM-BERT Large	**0.81**
three-class	BiLSTM	0.56
	BiCLSTM	0.53
	TACAM-WE	0.54
	TACAM-KG	0.56
	CAM-BERT Base	0.65
	TACAM-BERT Base	0.66
	CAM-BERT Large	0.67
	TACAM-BERT Large	**0.69**

Table 1: In-Topic

5.4 Cross-Topic Results

Our cross-topic results are presented in Table 2. In this experiment, which reflects a real-life argument search scenario, we want to prove our two hypotheses:

- When classifying potential arguments, it is advantageous to take information about the topic into account.
- The context of an argument and topic context are important for the classification decision.

On the whole, we can see that our two hypotheses are confirmed. In the two-classes scenario the recurrent model improves if topic information is provided by knowledge graph embeddings. By using attention-based models with pre-trained weights we can observe a significant performance boost of eleven score points in average when considering topic information. However, the same model without topic information performs only slightly better than the recurrent models. Therefore, we conclude that both, topic information together with contexts of topic and argument, are important for the correct decision about a potential argument. We observe similar effects in the three-classes scenario. Although in average different contexts for the recurrent model have a similar effect, we can clearly observe that taking topic information into account improves classification results by one score points. The combination of transfer learning for context and topic information again outperforms all other approaches by far. At the same time, the pre-trained model without topic information achieves a macro-f1 score of 0.61 which is 3 points lower than with topic information.

5.5 Topic Dependent Cross-Topic Results

As was shown in the previous subsection, argument classification produces satisfying results, especially if topic information and contexts are taken into account. In this set of experiments we evaluate the ability of different models to classify dependent on the topic. Therefore, a sentence may be considered to be an argument for one topic but be non argumentative for another. We argue that this is important, especially if text corpora are large, to filter out argumentative candidates which are arguments for different topics. To evaluate the models ability to perform well in topic dependent classification we extend our dataset and change the experimental setting. For each topic we select a number of related terms. These are words which come from a similar context as a topic but it is very unlikely that the topic's argument are valid arguments for them. The list of related terms for each topic is provided in Table 3. For 50% of argumentative sentences selected randomly from the test set, we replace the topic by one of the related terms of the topic and change the sentence label in the test set to non-argumentative. Therefore, to perform well on this task, a model should be able to recognize argumentative sentences in the context of the topic. To train for this task we correspondingly augment the training data. We keep the original training data and additionally select 50% of argumentative sentences from the training set, select one of the related terms as topic, label them as non-argumentative and insert them into the training set. For this task we compare our model, which performed best on the original cross-topic task and compare it with the state-of-the-art approach BiCLSTM . We also include the same models without topic information to see, whether topic information is still helpful or if the models get confused instead.

The results for topic dependent classification are presented in Table 4. For the two-classes problem we observe a massive performance drop of ten points in macro-f1 score for the BiCLSTM model. Nonetheless, the model still makes use of topic information and outperforms the standard BiLSTM by two macro-f1 score points. Our approach TACAM-BERT Base is more robust, the performance falls by moderate four score points and the gap to the counterpart model without topic information is incredible 17 score points large. We observe a similar behaviour in the three-classes scenario. Our TACAM-BERT Base approach achieves the same average score as in the original cross topic task. In contrast the performance of the BiCLSTM model drops by 11 score points and it even performs worse than the same model without topic information on this more complex task. Thus we conclude that unlike previous models our approaches are indeed able to grasp the context of the argument and topic and are able to relate them with each other.

6 CONCLUSION

In this paper, we introduce a new approach for argument mining which takes a topic of the potential argument into account. We

	Method	Topics								
		Abortion	Cloning	Death penalty	Gun control	Marij. legal.	Min. wage	Nucl. energy	School unif.	⌀
two-classes	BiLSTM	0.61	0.72	0.70	0.75	0.64	0.62	0.67	0.54	0.66
	BiCLSTM	0.67	0.71	0.71	0.73	0.69	0.75	0.71	0.58	0.70
	TACAM-WE	0.64	0.71	0.70	0.74	0.64	0.63	0.68	0.55	0.66
	TACAM-KG	0.62	0.69	0.70	0.75	0.64	0.76	0.71	0.56	0.68
	CAM-BERT Base	0.61	0.77	0.74	0.76	0.74	0.61	0.76	0.73	0.72
	CAM-BERT Large	0.62	0.79	0.75	0.77	0.77	0.65	0.75	0.73	0.73
	TACAM-BERT Base	0.78	0.77	**0.78**	0.80	**0.79**	0.83	0.80	**0.83**	0.80
	TACAM-BERT Large	**0.79**	**0.78**	**0.78**	**0.81**	**0.79**	**0.84**	**0.83**	0.82	**0.80**
three-classes	BiLSTM	0.47	0.52	0.48	0.48	0.44	0.42	0.48	0.42	0.46
	BiCLSTM	0.49	0.52	0.46	0.51	0.46	0.44	0.47	0.42	0.47
	TACAM-WE	0.47	0.52	0.47	0.48	0.46	0.46	0.48	0.41	0,47
	TACAM-KG	0.46	0.51	0.47	0.47	0.46	0.48	0.47	0.41	0.47
	CAM-BERT Base	0.38	0.63	0.53	0.49	0.54	0.54	0.61	0,50	0.53
	TACAM-BERT Base	0.42	0.68	0.54	0.50	0.60	0.49	0.64	**0.69**	0.57
	CAM-BERT Large	0.53	0.67	0.56	0.53	0.59	0.66	0.67	0.66	0.61
	TACAM-BERT Large	**0.54**	**0.69**	**0.59**	**0.55**	**0.63**	**0.69**	**0.71**	**0.69**	**0.64**

Table 2: Cross-Topic

Topic	Related terms				
abortion	euthanasia	teenage pregnancy	family	medical procedure	rape
cloning	biology	species	religion	organ donation	modified food
death penalty	politics	ethic	prison	homicide	sentence
gun control	safety	school shooting	robbery	regulation	police state
marijuana legalization	drugs	medicine	relaxation	freedom	liberty
minimum wage	social justice	slavery	automation	economic crisis	stagnation
nuclear energy	environment	employment	industry	pollution	climate change
school uniforms	equality	social justice	individualism	clothing	mobbing

Table 3: Related terms for each topic

	Method	Topics								
		Abortion	Cloning	Death penalty	Gun control	Marij. legal.	Min. wage	Nucl. energy	School unif.	⌀
two-classes	BiLSTM	0.57	0.59	0.53	0.59	0.62	0.62	0.59	0.57	0.58
	BiCLSTM	0.62	0.72	0.46	0.46	0.76	0.60	0.69	0.45	0.60
	CAM-BERT Base	0.56	0.63	0.60	0.62	0.61	0.55	0.60	0.53	0.59
	TACAM-BERT Base	**0.68**	**0.77**	**0.78**	**0.79**	**0.82**	**0.85**	**0.79**	**0.58**	**0.76**
three-classes	BiLSTM	0.39	0.39	0.37	0.36	0.39	0.42	0.40	0.39	0.39
	BiCLSTM	**0.46**	0.34	0.29	0.35	0.42	0.29	0.47	0.30	0.36
	CAM-BERT Base	0.42	0.50	0.42	0.42	0.48	0.51	0.50	0.49	0.47
	TACAM-BERT Base	0.44	**0.60**	**0.52**	**0.49**	**0.61**	**0.65**	**0.62**	**0.55**	**0.56**

Table 4: Topic dependent cross-topic classification results

hypothesize that considering information about the topic of a potential argument and their contexts should lead to better argument recognition. We present multiple ways to include topic and contexts into the argument mining process. Precisely, we show how contexts from word embeddings, Knowledge Graph embeddings and models pre-trained on other tasks can be integrated into our approach. Our experimental results clearly show that considering topics in the decision process leads to better results in almost all considered cases. Especially our approach with topic information in connection with context from pre-trained models improves state-of-the-art approach by far in the real-world scenario. We also show that in contrast to current state-of-the-art methods, our approach is robust and able to perfectly grasp the context of topic and potential argument. For future work we plan to focus more on Knowledge Graphs and other external context sources. In detail, we want to use information gathered from knowledge graphs not only for topics

but also on the argument side. We also plan to investigate different Knowledge Graph embedding techniques and combine different Knowledge Graphs in the same model. For instance, a combination of fact based knowledge graphs like DBPedia [16] and Wikidata [41] with knowledge graphs like WordNet [20] and FrameNet [1, 2] which focus on lexical similarities could further increase the representation quality of the context. Additional datasets with topic information about more topics will also deepen our understanding of the interplay between context and arguments and potentially further increase the performance of the argumentation models.

ACKNOWLEDGMENTS

This work has been funded by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A. This work has also been funded by the Deutsche Forschungsgemeinschaft (DFG) within the project Relational Machine Learning for Argument Validation (ReMLAV), Grant Number SE 1039/10-1, as part of the Priority Program "Robust Argumentation Machines (RATIO)" (SPP-1999). The authors of this work take full responsibilities for its content.

REFERENCES

[1] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of the 17th International Conference on Computational Linguistics - Volume 1 (COLING '98)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 86–90. https://doi.org/10.3115/980451.980860

[2] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1 (ACL '98/COLING '98)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 86–90. https://doi.org/10.3115/980845.980860

[3] Andrzej Białecki, Robert Muir, Grant Ingersoll, and Lucid Imagination. 2012. Apache lucene 4. In *SIGIR 2012 workshop on open source information retrieval*. 17.

[4] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146.

[5] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. In *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 2787–2795. http://papers.nips.cc/paper/5071-translating-embeddings-for-modeling-multi-relational-data.pdf

[6] Andrew M Dai and Quoc V Le. 2015. Semi-supervised sequence learning. In *Advances in neural information processing systems*. 3079–3087.

[7] Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the Essence of a Claim? Cross-Domain Claim Identification. *CoRR* abs/1704.07203 (2017).

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018).

[9] Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural End-to-End Learning for Computational Argumentation Mining. *CoRR* abs/1704.06104 (2017).

[10] James B. Freeman. 2011. *Argument Structure: Representation and Theory*. Springer.

[11] Shalini Ghosh, Oriol Vinyals, Brian Strope, Scott Roy, Tom Dean, and Larry P. Heck. 2016. Contextual LSTM (CLSTM) models for Large scale NLP tasks. *CoRR* abs/1602.06291 (2016). arXiv:1602.06291 http://arxiv.org/abs/1602.06291

[12] Ivan Habernal and Iryna Gurevych. 2016. Argumentation Mining in User-Generated Web Discourse. *CoRR* abs/1601.02403 (2016).

[13] Ivan Habernal, Maria Sukhareva, Fiana Raiber, Anna Shtok, Oren Kurland, Hadar Ronen, Judit Bar-Ilan, and Iryna Gurevych. 2016. New Collection Announcement: Focused Retrieval Over the Web. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16)*. ACM, New York, NY, USA, 701–704. https://doi.org/10.1145/2911451.2914682

[14] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (Nov. 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[15] Xinyu Hua and Lu Wang. 2017. Understanding and Detecting Supporting Arguments of Diverse Types. (04 2017).

[16] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2015. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal* 6, 2 (2015), 167–195. http://jens-lehmann.org/files/2015/swj_dbpedia.pdf

[17] Marco Lippi and Paolo Torroni. 2015. Argument Mining: A Machine Learning Perspective. In *Theory and Applications of Formal Argumentation*, Elizabeth Black, Sanjay Modgil, and Nir Oren (Eds.). Springer International Publishing, Cham, 163–176.

[18] Marco Lippi and Paolo Torroni. 2015. Context-independent Claim Detection for Argument Mining. In *Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI'15)*. AAAI Press, 185–191. http://dl.acm.org/citation.cfm?id=2832249.2832275

[19] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[20] George A. Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM* 38, 11 (Nov. 1995), 39–41. https://doi.org/10.1145/219717.219748

[21] Huy Nguyen and Diane Litman. 2015. Extracting Argument and Domain Words for Identifying Argument Components in Texts. In *Proceedings of the 2nd Workshop on Argumentation Mining*. Association for Computational Linguistics, Denver, CO, 22–28. https://doi.org/10.3115/v1/W15-0503

[22] Huy Nguyen and Diane Litman. 2016. Context-aware Argumentative Relation Mining. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 1127–1137. https://doi.org/10.18653/v1/P16-1107

[23] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. 2015. A review of relational machine learning for knowledge graphs. *Proc. IEEE* 104, 1 (2015), 11–33.

[24] Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation Mining: The Detection, Classification and Structure of Arguments in Text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law (ICAIL '09)*. ACM, New York, NY, USA, 98–107. https://doi.org/10.1145/1568234.1568246

[25] Andreas Peldszus and Manfred Stede. 2013. From Argument Diagrams to Argumentation Mining in Texts: A Survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)* 7, 1 (January 2013), 1–31. https://ideas.repec.org/a/igg/jcini0/v7y2013i1p1-31.html

[26] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[27] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

[28] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf* (2018).

[29] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1 (2019), 8.

[30] Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. 2018. ArgumenText: Searching for Arguments in Heterogeneous Sources. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. Association for Computational Linguistics, 21–25. https://doi.org/10.18653/v1/N18-5005

[31] Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. 2018. ArgumenText: Searching for Arguments in Heterogeneous Sources. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. Association for Computational Linguistics, New Orleans, Louisiana, 21–25. https://doi.org/10.18653/v1/N18-5005

[32] Christian Stab and Iryna Gurevych. 2014. Identifying Argumentative Discourse Structures in Persuasive Essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. 46–56. http://aclweb.org/anthology/D/D14/D14-1006.pdf

[33] Christian Stab and Iryna Gurevych. 2017. Parsing Argumentation Structures in Persuasive Essays. *Computational Linguistics* 43, 3 (Sept. 2017), 619–659. https://doi.org/10.1162/COLI_a_00295

[34] Christian Stab, Tristan Miller, and Iryna Gurevych. 2018. Cross-topic Argument Mining from Heterogeneous Sources Using Attention-based Neural Networks. *CoRR* abs/1802.05758 (2018).

[35] Ola Svenson. 1979. Process descriptions of decision making. *Organizational Behavior and Human Performance* 23, 1 (1979), 86 – 112. https://doi.org/10.1016/0030-5073(79)90048-5

[36] Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument Mining: Extracting Arguments from Online Dialogue. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Prague, Czech Republic, 217–226. https://doi.org/10.18653/v1/W15-4631

[37] Stephen E. Toulmin. 1958. *The Uses of Argument*. Cambridge University Press.

[38] Dietrich Trautmann, Johannes Daxenberger, Christian Stab, Hinrich Schütze, and Iryna Gurevych. 2019. Robust Argument Unit Recognition and Classification. *CoRR* abs/1904.09688 (2019).

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[40] Maria Paz Garcia Villalba and Patrick Saint-Dizier. 2012. Some Facets of Argument Mining for Opinion Analysis. In *COMMA*.

[41] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM* 57, 10 (Sept. 2014), 78–85. https://doi.org/10.1145/2629489

[42] Henning Wachsmuth, Martin Potthast, Khalid Al Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017. Building an Argument Search Engine for the Web. In *Proceedings of the 4th Workshop on Argument Mining*. Association for Computational Linguistics, Copenhagen, Denmark, 49–59. https://doi.org/10.18653/v1/W17-5106

[43] Douglas Walton. 2012. Argument Mining by Applying Argumentation Schemes. *Studies in Logic* 4 (04 2012).

[44] Adam Wyner, Raquel Mochales-Palau, Marie-Francine Moens, and David Milward. 2010. *Approaches to Text Mining Arguments from Legal Cases*. Springer Berlin Heidelberg, Berlin, Heidelberg, 60–79. https://doi.org/10.1007/978-3-642-12837-0_4

3 Argument Mining Driven Analysis of Peer-Reviews

The chapter includes the following publication:

> **Michael Fromm**, Evgeniy Faerman, Max Berrendorf, Siddharth Bhargava, Ruoxia Qi, Yao Zhang, Lukas Dennert, Sophia Selle, Yang Mao, and Thomas Seidl. „Argument Mining Driven Analysis of Peer-Reviews.“ In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35 (May 2021), pp. 4758–4766. URL: `https://ojs.aaai.org/index.php/AAAI/article/view/16607`

and the code is available at:

> `https://github.com/fromm-m/aaai2021-am-peer-reviews`

Declaration of Authorship Michael Fromm, Evgeniy Faerman, and Max Berrendorf developed and conceptualized the research idea. Siddharth Bhargava, Ruoxia Qi, Yao Zhang, Lukas Dennert, Sophia Selle, and Yang Mao implemented and evaluated the approach in the practical project. Michael Fromm, Evgeniy Faerman, and Max Berrendorf analyzed the results and discussed the findings. Michael Fromm, Evgeniy Faerman and Max Berrendorf wrote the manuscript. All authors revised the manuscript.

Argument Mining Driven Analysis of Peer-Reviews

Michael Fromm,[1] **Evgeniy Faerman,**[1] **Max Berrendorf,**[1] **Siddharth Bhargava,**[2] **Ruoxia Qi,**[2] **Yao Zhang,**[2] **Lukas Dennert,**[2] **Sophia Selle,**[2] **Yang Mao**[2] and **Thomas Seidl**[1]

[1] Database Systems and Data Mining, LMU Munich, Germany
[2] LMU Munich, Germany
{fromm, faerman, berrendorf}@dbs.ifi.lmu.de

Abstract

Peer reviewing is a central process in modern research and essential for ensuring high quality and reliability of published work. At the same time, it is a time-consuming process and increasing interest in emerging fields often results in a high review workload, especially for senior researchers in this area. How to cope with this problem is an open question and it is vividly discussed across all major conferences. In this work, we propose an Argument Mining based approach for the assistance of editors, meta-reviewers, and reviewers. We demonstrate that the decision process in the field of scientific publications is driven by arguments and automatic argument identification is helpful in various use-cases. One of our findings is that arguments used in the peer-review process differ from arguments in other domains making the transfer of pretrained models difficult. Therefore, we provide the community with a new peer-review dataset from different computer science conferences with annotated arguments. In our extensive empirical evaluation, we show that Argument Mining can be used to efficiently extract the most relevant parts from reviews, which are paramount for the publication decision. The process remains interpretable since the extracted arguments can be highlighted in a review without detaching them from their context.

Introduction

Argumentation is a process of bringing together and organizing reasons to convince a reasonable critic to accept or refuse a certain standpoint (Van Eemeren, Grootendorst, and van Eemeren 2004). It is an essential part of each rational decision-making process and after the decision is made, argumentation is important for its explanation and justification (Amgoud and Prade 2009). An important step in the argumentation process is the identification of arguments. Generally speaking, there is a difference between *argumentative* and *informative* content: Argumentative content expresses evidence or reasoning used to either oppose or support a given point. Informative parts often contain background information and describe how entities appear and act in the world.

In the last years, *Argument Mining* (AM) approaches have been applied in many fields and for different types of texts, such as encyclopedic articles (Aharoni et al. 2014), student essays (Stab and Gurevych 2014b), web discourse (Habernal and Gurevych 2016) or political speeches (Haddadan, Cabrio, and Villata 2019). AM techniques build the backbone of an IBM AI system *Project Debater*, which has the ambitious goal to debate humans on complex topics. This work aims to further extend the application of AM to the novel domain of scientific *peer reviews*. Peer reviewing is a cornerstone of today's academic editorial decision-making process in nearly all scientific disciplines. The peer-reviewers, who are usually not part of the editorial team, are experts in the corresponding research field and their task is the critical evaluation of the work proposed for publication. We argue that peer-reviewing can also be seen as an argumentation process, where the reviewers make up their minds about the examined publications and try to convince the editorial team by providing arguments in favor of or against acceptance. While the evaluation or review usually comprises different parts, such as a summary of the work or additional background information about the topic, the reviewers' pro and contra arguments are often the most relevant for making the final decision. Consequently, we envision that the automatic identification of argumentative content can improve and simplify different peer-review process phases. One possible use-case is to provide editors or meta-reviewers, coresponsible for the final decision, with an overview of arguments from all reviews and let them focus on the most relevant ones. For instance, after reading only the highlighted arguments in Figure 1, it is possible to get a good idea about the paper's strong and weak points. Another possible use-case is to support the reviewers by providing information about (missing) argumentation. For example, the author of the review in Figure 1 provides a detailed description of the empirical evaluation, but it is not completely clear from the text whether the reviewer is satisfied with the proposed evaluation criteria.

In this paper, we propose the application of AM to the domain of peer-reviewing. To this end, we collect a new dataset containing peer-reviews from different computer science conferences. We define a suitable AM annotation schema and annotate the dataset. We investigate the applicability of state-of-the-art AM techniques in an extensive empirical evaluation. Among others, we study the transferability of models trained on data from different domains to our task

Example Review

Summary: As the title suggests the paper focusses mostly on a negative result: Mutual information (MI) estimators obtained by variational methods have severe limitations that make them potentially not useful for down stream tasks. Besides highlighting the problems with variational MI estimators the authors suggest a modification to slightly improve the performance of MI estimators based on partition functions by reducing their variance when MI is high. **The authors give a good overview / introduction of various approaches to variational MI estimation by discriminative and generative methods.** Generally, MI estimation involves the estimation of the KL divergence between the joint distribution and the product of the marginals. The authors present a unifying view on the different approaches that optimizes the log density ratio required for the KL divergence over the space of log density ratios. Discriminative approaches model the density ratio directly (through e.g. neural network models) and generative approaches model the separate densities (as generative models where it is possible to evaluate the (conditional) probabilities / likelihoods of the data generating process). The authors prove that discriminative approaches that are based on the partition function approach suffer from high variance where mutual information is high (Theorem 2). The estimator based on a finite sample has high variance even if the density ratio approximation is correct. (The partition function approach is a way of staying constrained to the log density ratio function space.) This high variance problem is something that has previously been observed empirically and is the main theoretical point that is being made about limitations of MI estimators. In order to slightly alleviate the problem of high variance the authors suggest a way of biasing MI estimators by clipping the density ratio estimates through a constant chosen as a hyper-parameter. They prove that their clipping approach reduces variance and therefore introduces a bias variance tradeoff. In their later experiments the clipped version of the discriminative approach performs much better in terms of variance than without clipping and also better than a generative approach. In order to empirically evaluate the quality of MI estimators the authors suggest three criteria that they call self-consistency: (i) independence, (ii) data processing, (iii) additivity Self-consistency is evaluated experimentally on images where mutual information is computed between original image and image with part covered. The authors claim and experimentally show that discriminative approaches fail in (iii) and generative approaches fail in (i), (ii). Overall, variational MI approaches do not satisfy self-consistency. Evaluation: I suggest to accept the paper. **The theoretical contribution of showing the variance limitation of discriminative approaches seems significant.** That insight leads to the idea that clipping can be a useful bias that significantly reduces variance without making the already biased anyways results much worst in the experiments. However, I also feel like - *the paper is not yet as focused as it could be. It contains many concepts that could need a little bit more space.*

- Suggestions:

- Page 2: Nitpick, but in the definition of pseudo-formula using pseudo-formula twice is not super readable on the first read

- Page 2: In the definition of pseudo-formula clearify whether pseudo-formula is a marginal or a joint density (as pseudo-formula is the cumulative joint)

Figure 1: Example review for an ICLR'20 submission with labeling: Arguments in favor of acceptance are shown in green; red denotes arguments against it.

and the generalization across different conferences. Furthermore, we empirically validate our assumption about the importance of arguments for the decision-making process in academic publishing.

Related Work

Argument Mining

Argument Mining (AM) is the task of recognizing argument components (Palau and Moens 2009; Habernal and Gurevych 2016; Stab and Gurevych 2017; Hua and Wang 2017; Nguyen and Litman 2015) and their relations (Stab and Gurevych 2017; Nguyen and Litman 2016). The basis of AM are argumentation schemes that define the structure of the argument components and the relations between them. There is no universally accepted theory of argumentation (Van Eemeren, Grootendorst, and Kruiger 2019), and over time, argumentation schemes of varying complexity have been suggested in the literature (Toulmin 1958; Walton 2012; Freeman 2011; Stab and Gurevych 2014b). The original model by Toulmin (1958) comprises *claims* as an assertion for general acceptance, *data* (also often called premises) as the source of evidence to establish the claim, a *warrant* to justify the inference from a premise to a claim, *backing* (facts behind the warrant), a *qualifier* (degree of certainty for the inference) and *rebuttals*. The model has often been adopted in literature and most of the time, only premises and claims are used as argument components. However, it was observed that arguments in many text types have a more straightforward structure, e.g., models trained on a single dataset to identify *claims* do not generalize well to other document types (Daxenberger et al. 2017). Furthermore, annotating a dataset crawled from heterogeneous text sources leads to a low agreement among annotators (Habernal and Gurevych 2016; Miller, Sukhareva, and Gurevych 2019). Also, specific argument components (backing, warrant) appearing in the Toulmin-Scheme (Toulmin 1958) are often stated implicitly (van Eemeren et al. 2003; Habernal and Gurevych 2016). An argumentative scheme recently proposed by Stab, Miller, and Gurevych (2018) omits these components and simply distinguishes between (*supporting*/*opposing*) arguments and non-argumentative text parts. Its reasonableness is confirmed on the one hand by relatively high agreement among reviewers, and on the other hand by the model performance on texts from heterogeneous sources, see e.g. (Fromm, Faerman, and Seidl 2019). Furthermore, it was observed that the distinction between *supporting* and *opposing* arguments is more challenging than the distinction between argumentative and non-argumentative parts (Trautmann et al. 2020b,a; Fromm, Faerman, and Seidl 2019).

The development of models for the identification of argument components according to an argumentative scheme is similar to other NLP disciplines. Previous approaches rely on feature engineering (Habernal and Gurevych 2016; Lawrence and Reed 2015; Stab and Gurevych 2014a), more recent methods apply neural networks models. Guggilla, Miller, and Gurevych (2016) were the first to apply recurrent neural networks for AM. The state-of-the-art performance

in AM is achieved with pre-trained transformer-based architectures (Fromm, Faerman, and Seidl 2019; Trautmann et al. 2020a; Reimers et al. 2019).

A popular real-life application of AM techniques are argument search engines such as *argumenText*[1] (Stab et al. 2018) and *args*[2] (Wachsmuth et al. 2017) which allow argument retrieval according to a user-defined topic. AM is applied in the preprocessing step, where arguments are extracted from documents before they are indexed by a search engine.

Application of NLP for Peer-reviewing Process

So far, AM for scientific peer-reviews has received little attention. Hua et al. (2019) introduce a dataset with propositions in scientific reviews. The annotation schema is comprised of components that often appear in reviews such as *requests, facts, evaluations* or *quotes*. The dataset is annotated on a sentence level and the main focus is to study the usage of different propositions across venues. In our application, we are interested in arguments directly affecting the decision process, and therefore, the stance of the argument bears essential information. Since this information is missing in Hua et al. (2019), this annotation schema is not suitable for our application. Closely related is Xiao et al. (2020) work, where the goal is to automatically detect the problem description in peer-reviews. However, although the problems can also be considered opposing arguments, it is crucial to consider both positive and negative arguments for our application.

Other related works deal with different aspects of the peer-reviewing process. In Plank and van Dalen (2019), the authors introduce a dataset with scientific reviews and analyze it based on the title, abstract, and review text on how well the citation impact of a paper can be predicted. Gao et al. (2019) study the effect of author replies in the rebuttal phase. *Argumentative zoning* (Teufel, Siddharthan, and Batchelor 2009) analyzes the rhetorical and argumentative structure of scientific papers with intending to convince reviewers that the knowledge claim of the paper is valid.

Dataset

We use the OpenReview[3] platform and the OpenReview-Crawler[4] to retrieve peer-reviews. We collect all reviews from six computer-science conferences listed in Table 1. The annotated dataset [5] and the code [6] is available.

There, we additionally provide basic statistics about conferences and collected reviews.

Preprocessing

In a first preprocessing step, we replace URLs, escape sequences, encapsulated mathematical formulas, Unicode symbols and markdown with a corresponding type placeholder token respectively, e.g. `<URL>` for URLs. Furthermore, we remove multiple consecutive whitespaces and split review texts into sentences using the `PunktSentenceTokenizer` from NLTK.[7] To further improve the sentence splitting results, we provide the tokenizer with a set of idioms and abbreviations commonly used in scientific texts to avoid sentence splitting in the middle or after them.[8] Finally, we remove all sentences with less than three tokens and go through the dataset manually and remove non-interpretable sentences.

From 12,135 collected reviews, we sample 77 for the annotation. To this end, we first sample a conference uniformly at random and then a review from the conference.[9] We use stratified sampling to ensure that sampled reviews reflect the following three characteristics of original review distribution for each conference: Review-Rating (1-4), Paper-Decision (acceptance / rejection), and Review-Length.

Annotation

Scheme We use a simple argumentation scheme proposed in Stab, Miller, and Gurevych (2018), which distinguishes between non-arguments, supporting arguments and attacking arguments, which we denote as `NON/PRO/CON` accordingly. While this simple scheme grasps argumentative context, the annotation is easier since annotators are not required to consider complex relationships between argumentative components. Furthermore, it is also flexible enough to capture argumentative parts that are not attributable to the single argument type. For instance, in our dataset, we often observe rhetorical questions that criticize the paper's vagueness under review. The annotation scheme can also be interpreted as a flat version of the *claim-premise* model: There is a single *claim*, *"The paper should be accepted"*, and arguments are premises that either attack or support the claim.

Annotation Process In total, we have seven annotators, all of whom are graduate-level computer science students. The annotation is made token-wise and when presented a review, an annotator chooses argumentative text spans and assigns labels with the argument type to it. The document parts which are not explicitly annotated are considered to be non-argumentative. We refer to this annotation as *token-level* annotation.

Each review is randomly assigned to three different annotators. We resolve situations when a token is assigned with different labels by different annotators with a majority vote. In case a token is assigned with three different labels, we ask a independent fourth annotator who did not previously annotate the review to make the final annotation decision.

To obtain *sentence-level* annotations from annotated tokens, we mainly follow the procedure described in Trautmann et al. (2020a). Sentences without argumentative tokens are annotated with the label `NON`. For sentences con-

[1] www.argumentsearch.com

[2] www.args.me

[3] https://openreview.net/

[4] https://openreview-py.readthedocs.io/en/latest/getting_data.html

[5] https://zenodo.org/record/4314390

[6] https://github.com/fromm-m/aaai2021-am-peer-reviews

[7] https://www.nltk.org/

[8] The manually defined set contains e.g. "e.g", "i.e.", "et al.", "Fig.", etc.

[9] We end up with 15 reviews for iclr20, 14 reviews for iclr19 and 12 per each other conference

Conference	Number of Papers	Number of Reviews	Acceptance rate	avg words
ICLR'19	1,419	4,332	35 %	403
ICLR'20	2,213	6,722	27 %	409
MIDL'19	59	178	80 %	362
MIDL'20	144	544	55 %	255
NeuroAI'19	62	174	68 %	305
GI'20	65	174	82 %	507
Total	3,962	12,135	-	368

Table 1: Dataset statistics

	PRO	CON	NON	Total
number of tokens	3,259 (12%)	10,559 (34%)	14,684 (54%)	28,502
number of sentences	203 (14%)	640 (46%)	558 (40%)	1,401

Table 2: The table shows the distribution of the classes in the datasets. The distribution of the labels in the token-level dataset is skewed towards NON, and in the sentence-level dataset towards CON.

taining argumentative tokens, we count the number of argumentative segments, which overlap with it. An argumentative segment is comprised of a sequence of tokens with the same argumentative label without interruption. The sentence is assigned with the label of the majority of segments. If the number of segments with both labels is the same, we count the number of tokens with argumentative labels and assign the most frequent token label. As a result, we get 28,502 annotated tokens and 1,401 sentences. Table 2 presents the resulting class distribution.

Agreement

The agreement among annotators is an important criterion for the reliability of the annotation. Since our annotations are done on a token level and we have more than two annotators per review, we use the Krippendorff's alpha (Krippendorff et al. 2016) family of measures to assess the annotation quality. Each annotation can be seen as a set of annotated segments $(start, stop, label)$, where $start$ and $stop$ denote the segment's bounds and $label$ its class. We include all three classes for the computation of agreement.[10] Krippendorff's alpha now considers all pairs of overlapping segments and compares the expected and the observed disagreements in the annotations. For better comparability we follow recent related work (Trautmann et al. 2020a) and compute the following two variants: ${}_{cu}\alpha$ only considers the agreement in the label, while ${}_{u}\alpha$ additionally takes the length of the overlap into account. For both variants, the perfect agreement corresponds to the value of 1, the score for a random agreement is zero and negative values are possible if the agreement is worse than random. For our annotation, we obtain ${}_{u}\alpha = 0.568$ and ${}_{cu}\alpha = 0.861$, which is comparable to related work (Trautmann et al. 2020a).

Another possibility to assess the agreement is to compute the Macro F_1 metric for individual annotators. In terms of the Macro F_1 score, the quality of our annotations is better than of comparable datasets (Trautmann et al. 2020a; Reimers et al. 2019), see Human Performance in Table 3. Thus, we conclude that our annotation is reliable for further experiments.

[10]The score also accounts for imbalanced classes, see e.g. (Artstein and Poesio 2008).

Experimental Setup

In the following, we discuss our experimental setup. The description applies for both token-level and sentence-level evaluation unless noted otherwise.

Problem Setting Our goal is to identify *supporting* and *opposing* arguments in scientific peer-reviews and separate them from non-argumentative text. To get a detailed analysis of the models' performance and possible bottlenecks, we first decouple the problem of *argument identification* from *stance detection* and solve them separately. Afterward, we jointly solve both problems by a single model and obtain a model performance for our desired application. Therefore, we define the following tasks:

1. Argumentation Detection: A binary classification of whether a text span is an argument. The classes are denoted by ARG and NON, where ARG is the union of PRO and CON classes.
2. Stance Detection: A binary classification whether an argumentative text span is supporting or opposing the paper acceptance. The model is trained and evaluated only on argumentative PRO and CON text spans.
3. Joint Detection: A multi-class classification between the classes PRO, CON and NON, i.e. the combination of argumentation and stance detection.

Evaluation

We split our dataset sentence-wise 7:1:2 into training, validation and test sets stratified by class, i.e. keeping the same ratio among classes in all three subsets. The validation set is used for hyperparameter optimization and early stopping, whereas the test set is only used to evaluate the final model performance reported in the result section. We report the macro F_1 score. The F_1 is defined as the harmonic mean of

precision and recall and Macro F_1 is the mean over the class-individual scores. Since Macro F_1 weights classes equally independently of class' size, it is insensitive to the class imbalance problem. We train each model ten times with different random seeds and report the mean performance.[11] To check the significance of our results, we use a two-sided t-test with a significance level 1%.

Methods

Since transfer learning achieves state-of-the-art results for AM on different datasets (Reimers et al. 2019; Fromm, Faerman, and Seidl 2019; Trautmann et al. 2020a) we also apply it for our task. We employ a transformer (Vaswani et al. 2017) based BERT model (Devlin et al. 2019) with fine-tuning on different datasets. We include the following model variants in our evaluation:

Majority Baseline The majority baseline labels the instances with the most frequent class.

ArgBERT To assess the new dataset necessity, we evaluate the zero-shot learning performance of a BERT model fine-tuned on another AM dataset annotated on token and sentence level with the same scheme (Trautmann et al. 2020a). The other dataset comprises heterogeneous data found on the internet, and therefore, the resulting model is supposed to be universally applicable.

PeerBERT-ArgInit We initialize the model with the weights of ArgBERT and additionally fine-tune it on our new dataset. We hypothesize that the model can take advantage of the argumentative structure learned on another dataset.

PeerBERT Smaller BERT model with 110M parameters fine-tuned on our dataset (based on `bert-base-cased`).

PeerBERT-L Larger BERT model with 340M parameters fine-tuned on our dataset (based on `bert-large-cased`).

Human Performance An interesting experiment for assessing the applicability of the proposed solution is the comparison with the human performance on the task. To compute the human performance, we treat each annotator analogously to the model. Therefore, we compare labels produced by each annotator to the final annotations and compute the Macro F_1 score. The reported score is the mean among scores of all annotators.[12]

Training

We use a weighted cross-entropy loss to tackle the class imbalance problem, where the weight is given as the reciprocal of the number of samples of this class. The class weights are defined individually for each task and dataset. The models are trained using either `bert-base-cased` or `bert-large-cased`, with training batch size 100 for bert-base and 32 for bert-large. We use the AdamW optimizer with a learning rate of 10^{-5} for all models and early stopping with a patience of 3.

[11] To avoid the clutter, we provide the variance across the different runs in the appendix

[12] The resulting score should be seen as the upper bound for human performance since we use the same annotations for ground-truth.

Results

In this section, we present the results of our experiments, which we have designed to answer the following research questions:

1. How well does the automatic mining of arguments work for peer-reviews?
2. Can we transfer knowledge from pre-existing annotated argumentation datasets?
3. How well does the approach generalize across different conferences?
4. How relevant are arguments in the decision making process for scientific publications?

Automatic Mining of Arguments

The results for the three AM tasks and all methods are summarized in Table 3. Our most important observation is that automatic argument extraction performs close to human performance and can be relied upon in the peer-review domain. Surprisingly, the detection of the stance in the peer-review domain appears to be considerably easier than identifying arguments. For other datasets annotated with the same scheme, we observe an inverse effect, see Table 4. Although there is no explicit stance detection experiment in the other works, we can infer it from the inferior results of joint detection compared against the argument detection results.

When comparing our results to other datasets on the token level, we observe that our results are substantially better, with a difference of about 10 % points. A reason might be that we operate on a single domain while other datasets contain heterogeneous documents covering multiple domains. However, we observe a significant performance difference when comparing our results on sentence and token level. To identify the reasons, we analyze the label ambiguity within sentences in our dataset. We found out that 22% of sentences for the argumentation detection task and 23% of those for the stance detection task contain tokens annotated with both classes. Therefore, we conclude that while it is still possible to achieve acceptable performance on the sentence level, the difference to the token level is more evident in our dataset.

Finally, the experiment regarding knowledge transfer from another AM dataset reveals transfer difficulties. The zero-shot performance is better than the majority vote only on the simpler stance detection task, but it is clearly outperformed by the models directly trained on our dataset. The additional intermediate fine-tuning step on the other AM dataset does not bring significant improvement either compared to directly fine-tuning on our dataset, cf. PeerBERT.

Training Set Size Figure 2 presents the model performance for different training set sizes. We can observe that pretraining on the other AM dataset does not help, even if the training set is small. The performance saturates when about

Detection Level	Argument Sentence	Argument Token	Stance Sentence	Stance Token	Joint Sentence	Joint Token
Majority Baseline	0.351	0.350	0.423	0.434	0.234	0.233
ArgBERT	0.316	0.353	0.719	0.644	0.203	0.241
PeerBERT-ArgInit	0.718	0.877	0.852	0.862	0.734	0.796
PeerBERT	**0.789**	0.896	0.893	0.849	0.728	0.808
PeerBERT-L	0.763	**0.900**	**0.936**	**0.930**	**0.757**	**0.839**
Human Performance	0.885	0.873	0.978	0.980	0.881	0.860

Table 3: Overview of the results for different Argument Mining tasks on token and sentence level. We show results in terms of Macro F_1 for different BERT model variants, as well as the majority baseline and human performance estimate. In bold font, we highlight the best performance of our models per task and level.

Detection Level	Argument Sentence	Argument Token	Joint Sentence	Joint Token
UKP	0.810	-	0.690	-
AURC	-	0.782	0.725	0.743
Ours	0.789	0.900	0.757	0.839

Table 4: Comparison of maximum Macro F_1 values obtained for different datasets from literature, UKP (Stab, Miller, and Gurevych 2018; Fromm, Faerman, and Seidl 2019) and AURC (Trautmann et al. 2020a).

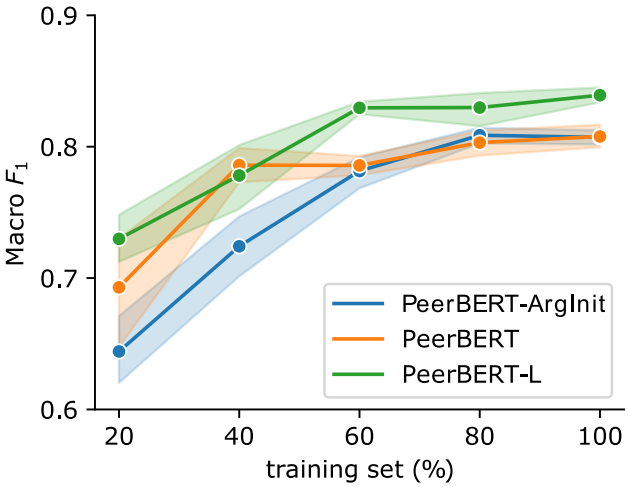

Figure 2: The Macro-F_1 evaluated on the task of joint prediction on the token level. The shaded areas indicate confidence intervals across ten runs with different random seeds.

Detection	Argument	Joint
ALL	0.891	0.823
NO-GI	0.873	0.791

Table 5: Comparison of Macro F_1 values for sentences from GI-20 reviews, when training with/without sentences from reviews from GI-20. All tasks are done on token-level.

60% of the training set is used. Therefore, we conclude that we have collected enough annotations. Similar behavior has been observed for the other tasks at both sentence and token level.

Generalization Across Conferences

In this section, we study the model's generalization to peer-reviews for papers from other (sub)domains. To this end, we reduce the test set to only contain reviews from the GI'20 conference. The focus of the GI'20 conference is Computer Graphics and Human-Computer Interaction, while the other conferences are focused on Representation Learning, AI and Medical Imaging. We consider the GI'20 as a subdomain since all conferences are from the domain of computer science. As a model, we choose our PeerBERT-L model and train on two different training sets:

NO-GI The original training dataset with all sentences from reviews of GI'20 removed.

ALL A resampling of the original training dataset of the same size as NO-GI, with sentences from all conferences.

Table 5 presents the experimental results. We observe a small performance decrease on both tasks, about two points on argument detection and three on joint detection tasks. At the same time, we also observe similar behavior when comparing results obtained on the whole test set (Table 3) and only on GI'20 reviews by the ALL model. Therefore the more considerable drop is not necessary due to the worse generalization and can be explained by the more challenging task. Overall, the drops are relatively small, and we conclude that the model generalizes well across subdomains.

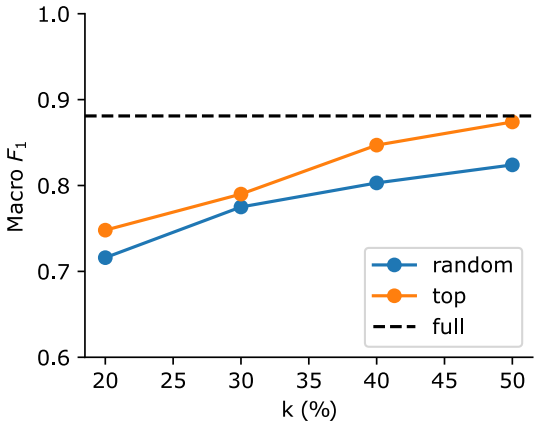

Figure 3: Evaluation of acceptance classification performance in F_1-measure based on different sentence selection methods. Using the top k% sentences according to argumentativeness likelihood results in superior performance compared to random selection. With 50% of the text, almost the same performance is reached as with the full review.

Relevance for Decision-making

In previous experiments, we have shown that peer-reviews contain arguments and these arguments can be identified automatically. In this section, we want to verify the usefulness of the extracted arguments for the decision making process. As a proxy to evaluate the usefulness, we design an experiment where the acceptance/rejection decisions made solely by considering arguments are compared to the decisions supported by taking full reviews into account. Therefore, we use the unannotated rest of our dataset and assign a probability to be an argument to each sentence with our best performing PeerBERT-L model. Now, we can compare three different settings for the decision-making process:

Full The decision-makers are allowed to see all reviews completely. This particularly includes decision suggestions often encountered in reviews that are not annotated as arguments in our dataset.

Top-K Arguments The decision-makers are only allowed to see the $k\%$ sentences with the highest probability to be arguments from each review. Note that the high probability to be identified as an argument does not necessarily correlate with the strength of the argument.

Random-K Decision-makers are only allowed to see $k\%$ randomly selected sentences from each review. We do not exclude explicit decision suggestions here.

We consider sentence level in this experiment despite the better performance of our model on the token-level. The main reason is a fair comparison with the Random-k setting, random sampling of words would result in large gaps and meaningless texts, especially for small k.

To avoid manual expenditure, we decide to apply a language model as a decision-maker. Since we also have a decision for each paper in our dataset, we train models to make an acceptance/rejection decision for the different settings described above. The standard BERT model is not directly applicable for this task since combining the reviews for a single paper often exceeds the input length restriction of at most 512 tokens. Therefore, we employ ToBERT (Pappagari et al. 2019), a model proposed for the classification of the long texts. It splits texts into multiple segments and individual segments are first used for the finetuning of the BERT model. In a second step, a second transformer model on the top combines representations of the segments and makes the final decision.

The results in terms of F_1-measure are given in Figure 3. We observe that selecting according to argumentativeness likelihood improves classification performance consistently in terms of F_1, compared to the random selection baseline, if at least a third of the review text is taken into consideration. The fraction of argumentative sentences in the annotated part of our dataset is 60%, cf. Table 2. We can achieve almost the same performance as the classifier trained on the full reviews while only considering 50% of the review. This is particularly impressive considering that reviews often already contain decision suggestions. Therefore, we conclude that arguments, which can be automatically extracted from reviews, are essential for the decision making process.

Conclusion

In this work, we have presented a new Argument Mining based approach for the assistance of different actors in the peer-review process. We have demonstrated that arguments are present in peer-reviews and that their identification with different stances can be made automatically. We have also shown that the peer-review domain is different from other previous Argument Mining applications, and therefore, there is a need for a new dataset. We have presented a new dataset that we make available for the community and have performed an extensive evaluation. We have also analyzed the editorial decision-making process and have empirically demonstrated that it is driven by argumentation.

In future work, we plan to address the problem of automatic determination of argument strength. Ranking arguments, according to their strength, is an undoubtedly useful feature for the potential application. For this purpose, we intend to extend our decision-making model and analyze single arguments' influence on the final decision.

Another useful feature, especially for the editorial team, would be identifying similar arguments in different reviews of the same paper.

Acknowledgments

This work has been funded by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A and by the Deutsche Forschungsgemeinschaft (DFG) within the project Relational Machine Learning for Argument Validation (ReMLAV), Grant Number SE 1039/10-1, as part of the Priority Program "Robust Argumentation Machines (RATIO)" (SPP-1999). The infrastructure for the course was provided by the Leibniz-Rechenzentrum. The authors of this work take full responsibilities for its content.

References

Aharoni, E.; Polnarov, A.; Lavee, T.; Hershcovich, D.; Levy, R.; Rinott, R.; Gutfreund, D.; and Slonim, N. 2014. A Benchmark Dataset for Automatic Detection of Claims and Evidence in the Context of Controversial Topics. In *Proceedings of the First Workshop on Argumentation Mining*, 64–68. Baltimore, Maryland: Association for Computational Linguistics. doi:10.3115/v1/W14-2109. URL https://www.aclweb.org/anthology/W14-2109.

Amgoud, L.; and Prade, H. 2009. Using arguments for making and explaining decisions. *Artificial Intelligence* 173(3-4): 413–436.

Artstein, R.; and Poesio, M. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34(4): 555–596.

Daxenberger, J.; Eger, S.; Habernal, I.; Stab, C.; and Gurevych, I. 2017. What is the Essence of a Claim? Cross-Domain Claim Identification. *CoRR* abs/1704.07203.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

Freeman, J. B. 2011. *Argument Structure: Representation and Theory*. Springer.

Fromm, M.; Faerman, E.; and Seidl, T. 2019. TACAM: Topic And Context Aware Argument Mining. In *2019 IEEE/WIC/ACM Int. Conf. on Web Intelligence, WI 2019, Thessaloniki, Greece, October 14-17, 2019*, 99–106.

Gao, Y.; Eger, S.; Kuznetsov, I.; Gurevych, I.; and Miyao, Y. 2019. Does My Rebuttal Matter? Insights from a Major NLP Conference. *CoRR* abs/1903.11367. URL http://arxiv.org/abs/1903.11367.

Guggilla, C.; Miller, T.; and Gurevych, I. 2016. CNN- and LSTM-based Claim Classification in Online User Comments. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2740–2751. Osaka, Japan: The COLING 2016 Organizing Committee. URL https://www.aclweb.org/anthology/C16-1258.

Habernal, I.; and Gurevych, I. 2016. Argumentation Mining in User-Generated Web Discourse. *CoRR* abs/1601.02403.

Haddadan, S.; Cabrio, E.; and Villata, S. 2019. Yes, we can! Mining Arguments in 50 Years of US Presidential Campaign Debates. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4684–4690. Florence, Italy: Association for Computational Linguistics. doi:10.18653/v1/P19-1463. URL https://www.aclweb.org/anthology/P19-1463.

Hua, X.; Nikolov, M.; Badugu, N.; and Wang, L. 2019. Argument Mining for Understanding Peer Reviews. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2131–2137. Minneapolis, Minnesota: Association for Computational Linguistics. doi:10.18653/v1/N19-1219. URL https://www.aclweb.org/anthology/N19-1219.

Hua, X.; and Wang, L. 2017. Understanding and Detecting Supporting Arguments of Diverse Types. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 203–208. Vancouver, Canada: Association for Computational Linguistics. doi:10.18653/v1/P17-2032. URL https://www.aclweb.org/anthology/P17-2032.

Krippendorff, K.; Mathet, Y.; Bouvry, S.; and Widlöcher, A. 2016. On the reliability of unitizing textual continua: Further developments. *Quality & Quantity* 50(6): 2347–2364.

Lawrence, J.; and Reed, C. 2015. Combining Argument Mining Techniques. In *Proceedings of the 2nd Workshop on Argumentation Mining*, 127–136. Denver, CO: Association for Computational Linguistics. doi:10.3115/v1/W15-0516. URL https://www.aclweb.org/anthology/W15-0516.

Miller, T.; Sukhareva, M.; and Gurevych, I. 2019. A Streamlined Method for Sourcing Discourse-level Argumentation Annotations from the Crowd. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 1790–1796. Minneapolis, Minnesota: Association for Computational Linguistics. doi:10.18653/v1/N19-1177. URL https://www.aclweb.org/anthology/N19-1177.

Nguyen, H.; and Litman, D. 2015. Extracting Argument and Domain Words for Identifying Argument Components in Texts. In *Proceedings of the 2nd Workshop on Argumentation Mining*, 22–28. Denver, CO: Association for Computational Linguistics.

Nguyen, H.; and Litman, D. 2016. Context-aware Argumentative Relation Mining. In *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1127–1137. Berlin, Germany: Association for Computational Linguistics.

Palau, R. M.; and Moens, M.-F. 2009. Argumentation Mining: The Detection, Classification and Structure of Arguments in Text. In *Proc. of the 12th Int. Conf. on Artificial Intelligence and Law*, ICAIL '09, 98–107. New York, NY, USA: ACM. ISBN 978-1-60558-597-0.

Pappagari, R.; Zelasko, P.; Villalba, J.; Carmiel, Y.; and Dehak, N. 2019. Hierarchical Transformers for Long Document Classification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 838–844. IEEE.

Plank, B.; and van Dalen, R. 2019. CiteTracked: A Longitudinal Dataset of Peer Reviews and Citations. In *BIRNDL@ SIGIR*, 116–122.

Reimers, N.; Schiller, B.; Beck, T.; Daxenberger, J.; Stab, C.; and Gurevych, I. 2019. Classification and Clustering of Arguments with Contextualized Word Embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 567–578. Florence, Italy:

Association for Computational Linguistics. doi:10.18653/v1/P19-1054. URL https://www.aclweb.org/anthology/P19-1054.

Stab, C.; Daxenberger, J.; Stahlhut, C.; Miller, T.; Schiller, B.; Tauchmann, C.; Eger, S.; and Gurevych, I. 2018. Argumentext: Searching for arguments in heterogeneous sources. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: demonstrations*, 21–25.

Stab, C.; and Gurevych, I. 2014a. Annotating Argument Components and Relations in Persuasive Essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 1501–1510. Dublin, Ireland: Dublin City University and Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C14-1142.

Stab, C.; and Gurevych, I. 2014b. Identifying Argumentative Discourse Structures in Persuasive Essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, 46–56. URL http://aclweb.org/anthology/D/D14/D14-1006.pdf.

Stab, C.; and Gurevych, I. 2017. Parsing Argumentation Structures in Persuasive Essays. *Computational Linguistics* 43(3): 619–659.

Stab, C.; Miller, T.; and Gurevych, I. 2018. Cross-topic Argument Mining from Heterogeneous Sources Using Attention-based Neural Networks. *CoRR* abs/1802.05758.

Teufel, S.; Siddharthan, A.; and Batchelor, C. 2009. Towards Domain-Independent Argumentative Zoning: Evidence from Chemistry and Computational Linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 1493–1502. Singapore: Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D09-1155.

Toulmin, S. E. 1958. *The Uses of Argument*. Cambridge University Press.

Trautmann, D.; Daxenberger, J.; Stab, C.; Schütze, H.; and Gurevych, I. 2020a. Fine-Grained Argument Unit Recognition and Classification. In *AAAI*.

Trautmann, D.; Fromm, M.; Tresp, V.; Seidl, T.; and Schütze, H. 2020b. Relational and Fine-Grained Argument Mining. *Datenbank-Spektrum* 1–7.

van Eemeren, F.; Blair, J.; Willard, C.; and Henkemans, A. 2003. *Anyone Who Has a View: Theoretical Contributions to the Study of Argumentation*, volume 8. Kluwer Academic Publishers. ISBN 978-1-4020-1456-7. doi:10.1007/978-94-007-1078-8.

Van Eemeren, F.; Grootendorst, R.; and van Eemeren, F. H. 2004. *A systematic theory of argumentation: The pragma-dialectical approach*. Cambridge University Press.

Van Eemeren, F. H.; Grootendorst, R.; and Kruiger, T. 2019. *Handbook of argumentation theory: A critical survey of classical backgrounds and modern studies*, volume 7. Walter de Gruyter GmbH & Co KG.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wachsmuth, H.; Potthast, M.; Al-Khatib, K.; Ajjour, Y.; Puschmann, J.; Qu, J.; Dorsch, J.; Morari, V.; Bevendorff, J.; and Stein, B. 2017. Building an Argument Search Engine for the Web. In *Proceedings of the 4th Workshop on Argument Mining*, 49–59. Copenhagen, Denmark: Association for Computational Linguistics. doi:10.18653/v1/W17-5106. URL https://www.aclweb.org/anthology/W17-5106.

Walton, D. 2012. Argument Mining by Applying Argumentation Schemes. *Studies in Logic* 4.

Xiao, Y.; Zingle, G.; Jia, Q.; Shah, H. R.; Zhang, Y.; Li, T.; Karovaliya, M.; Zhao, W.; Song, Y.; Ji, J.; et al. 2020. Detecting Problem Statements in Peer Assessments. In *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, 704–709.

4 Relational and Fine-Grained Argument Mining

The chapter includes following publication:

> Dietrich Trautmann, **Michael Fromm**, Volker Tresp, Thomas Seidl, and Hinrich Schütze. „Relational and Fine-Grained Argument Mining.“ In: *Datenbank-Spektrum* (2020), pp. 1–7

Declaration of Authorship The survey of our work was proposed, developed and conceptualized by Michael Fromm and Dietrich Trautmann. Michael Fromm, and Dietrich Trautmann wrote the manuscript and discussed it with Volker Tresp, Thomas Seidl, and Hinrich Schütze. All authors revised the manuscript.

Datenbank Spektrum (2020) 20:99–105
https://doi.org/10.1007/s13222-020-00341-z

SCHWERPUNKTBEITRAG

Relational and Fine-Grained Argument Mining

The LMU Munich project *ReMLAV* within the DFG Priority Program *RATIO "Robust Argumentation Machines"*

Dietrich Trautmann[1] · Michael Fromm[2] · Volker Tresp[2] · Thomas Seidl[2] · Hinrich Schütze[1]

Received: 5 February 2020 / Accepted: 29 May 2020 / Published online: 12 June 2020

Abstract
In our project *ReMLAV*, funded within the DFG Priority Program *RATIO* (http://www.spp-ratio.de/), we focus on relational and fine-grained argument mining. In this article, we first introduce the problems we address and then summarize related work. The main part of the article describes our research on argument mining, both coarse-grained and fine-grained methods, and on same-side stance classification, a relational approach to the problem of stance classification. We conclude with an outlook.

Keywords Argument Mining · Stance Classification · Relational Machine Learning

1 Introduction

In the project *ReMLAV*, funded within the DFG Priority Program *RATIO* (http://www.spp-ratio.de/), the *Center for Information and Language Processing (CIS)* and the *Chair for Database Systems and Data Mining (DBS)* at LMU Munich join forces to work on argument mining, an important problem in computational argumentation. Argument mining is the task of extracting argumentative sentences from large document collections to support argument search engines. We address two aspects of argument mining: argument extraction and stance classification.

✉ Dietrich Trautmann
dietrich@cis.lmu.de

Michael Fromm
fromm@dbs.ifi.lmu.de

Volker Tresp
tresp@dbs.ifi.lmu.de

Thomas Seidl
seidl@dbs.ifi.lmu.de

Hinrich Schütze
inquiries@cislmu.org

[1] Center for Information and Language Processing, LMU Munich, Munich, Germany

[2] Database Systems and Data Mining, LMU Munich, Munich, Germany

Argument extraction is the core task of argument mining by identifying those parts of a document that are argumentative. We address this problem on two levels, on the sentence-level (coarse-grained) and on the token-level (fine-grained). For sentence-level argument extraction (Sect. 3.1.1), our research focuses on representations that capture different types of information that can support this task. Sentences as a whole are classified as, e.g., argumentative vs. non-argumentative. For token-level argument extraction (Sect. 3.1.2), we formalize the problem as sequence labeling which is a novel argument mining approach. Each token in the document is labeled, e.g., as argumentative vs. non-ar-gumentative. Argumentative segments are then the set of tokens consisting of maximum sequences that are labeled as argumentative.

The second problem we address is stance classification, i.e., the classification of an argumentative segment or sen-

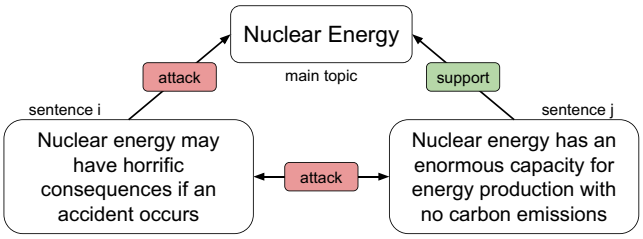

Fig. 1 Argumentative *sentences i* and *j* and the *main topic* [31], with *support* and *attack relations* between them

tences with either a PRO label (arguing for a topic or point of view) or with a CON label (arguing against the topic). One important concept in this context are argumentative relations. Fig. 1 shows examples for relations between argumentative sentences and the topic "nuclear energy". The relations are in this case supporting and attacking relations. Additionally, we develop methods to improve the overall stance classification with relational information, such as *same-side* and *not-same-side* in the same-side stance classification task (Sect. 3.2).

2 Related Work

2.1 Argumentation Schemes

A foundation for argument mining is an *argumentation sche-me*. An argumentation scheme defines what kind of arguments exist and the properties and relationships between them. Consequently, the main emphasis in argument mining lies in detecting argument components of argumentation schemes [12, 14, 16, 20, 27] and the relations between them [17, 27]. Different argumentation schemes of varying complexity have been suggested [8, 26, 30, 33].

However, many argument components (e.g., claims, prem-ises) do not generalize well across text types. Some works [6] show that it is not sufficient to train a single claim-detection model. Often the agreement between annotators during the dataset creation is low, since argumentation is a complex, highly subjective task [12]. Certain argument components (e.g., backing and warrant [30]) are often only implicitly stated [12]. Therefore, researchers have defined simpler and more tractable argumentation schemes.

In the simplest case, the argumentation scheme only differentiates between argumentative and non-argumentative text units. In a slightly more complex setting, stance information is also considered [28]. Computational argumentation models trained on these simpler argumentation schemes are often better applicable to a broader range of text genres. Based on these simpler schemes, two argument search engines, *ArgumenText*[1] [25] and *args*[2] [32] have been realized, where users can search a broad range of documents for certain topics.

Given the success of simpler argumentation schemes, we adopt them for our work.

2.2 Relational Machine Learning

A novel aspect of our approach is to model sets of arguments as graphs where each argument is a node and edges between arguments are relations like "attack" and "support", as shown in Fig. 1. This relational model allows us to make inferences about arguments in the context of related arguments, inferences that would not be possible if we looked at each argument in isolation.

Relational data is gaining in importance in machine learning. The literature review by Nickel et al. [18], with an emphasis on knowledge graph construction, discusses many current models and datasets for relational machine learning. One of the successful models presented is RESCAL [19], which is based on tensor factorization. This model works over triples of subject, predicate and object, with the predicate describing the relation between the subject and the object. This and similar models have been trained over large knowledge graphs such as YAGO [29], DBpedia [2] and Freebase [4]. This approach could conceivably also be applied to argument graphs, but this is not trivial. For example, subjects and objects in knowledge graphs generally occur in many different relations, but most arguments in text are unique if they are represented as sequences of words.

In this article, we adopt a simpler approach to relational information: we build a graph of arguments where known edges are either same-side (both PRO or both CON) or not-same-side (one is PRO, one is CON). By incorporating new arguments into this graph, we can infer their stance.

3 Argument Mining Tasks

For argument mining, a substantial text collection is required. Many large topic-specific textual corpora can readily be retrieved from the Internet. In addition, one can exploit Internet search engines to discover and download news or discussion documents. There are also crawled web data such as *Common Crawl*[3] that can be indexed with tools like *Elasticsearch*[4]. Other resources include the *Open Web Text* [11] corpus, which is based on documents (urls) submitted to the social media platform *Reddit*[5].

Argument mining models, which are trained on annotated datasets, can be applied on the previously mentioned corpora to extract argumentative sentences. The level of granularity varies in those models and two important ones are models that are trained on the sentence-level (coarse-grained) and on the token-level (fine-grained). In our approaches, the goal is to classify whether units (sentences or tokens) are supporting (PRO), attacking (CON) or neutral (NON) toward a controversial topic. Token-level models support extracting argumentative segments that are often

[1] www.argumentsearch.com.

[2] www.args.me.

[3] http://commoncrawl.org/.

[4] https://www.elastic.co/elasticsearch/.

[5] https://www.reddit.com/.

addressing only one specific aspect of larger arguments and thus can be more useful in further downstream applications. Fine-grained models also support capturing several segments with-in a sentence that address different aspects and have different stances.

Stance classification is of central importance in argument mining, e.g., in an argument search engine that gives the user PRO arguments on one side and CON arguments on the other. Stance classification is hard because it typically requires a lot of detailed world and background knowledge as well as larger context. We approach stance classification through same-side stance classification. Pairs of argumentative paragraphs, sentences or segments are classified as being on the same-side (same stance toward a topic) or not. The graph of all arguments (with same-side and non-same-side edges) is then exploited for more accurate stance classification.

3.1 Argument Extraction

3.1.1 Sentence-Level Models

In previous work [9], some of us addressed the problem of topic-focused argument extraction on the sentence-level. Examples of the type of sentences that we extract can be seen in Fig. 2 (lines 1-3). We define topic-focused argument extraction as argument extraction where a user-defined query topic (e.g., "nuclear energy") is given. The query topic is important for the argument extraction decision because a given sentence may be an argument supporting one topic, but not another. Since we cannot expect that available datasets cover all possible topics, the ability to generalize to unseen topics is an important requirement. Therefore, the better a machine learning model is capable of grasping the context of topic and of potential arguments, the better decisions it can make and the more confident it can be about its decisions. The work introduced recurrent and attention based networks that encode the topic information as an additional input besides the sentence. As context sources we relied on different external sources that provide the context information.

- Shallow **Word Embeddings** [3, 15, 21] are commonly used in natural-language-processing (NLP) applications and encode context information implicitly.
- **Knowledge Graphs** are heterogeneous multi-relational graphs that model information about the world explicitly. Information is represented as triples consisting of subject, predicate and object, where subject and object are entities and predicate stands for the relationship between them. Compared to textual data, knowledge graphs are structured, i.e., each entity and relationship has a distinct meaning, and the information about the modeled world is distilled in form of facts. These facts stem from texts, different databases, or are inserted manually. The reliability of these facts in (proprietary) knowledge graphs can be very high [18].
- Fine-tuning based **Transfer Learning** approaches [7, 23, 24] adapt whole models that were pre-trained on some (auxiliary) task to a new problem. This is different from feature-based approaches which provide pre-trained representations [5, 22] and require task-specific architectures for a new problem.

For the evaluation of our methods we used the UKP Sentential Argument Mining corpus [28]. It consists of more than 25,000 sentences from multiple text genres covering eight controversial topics. We have evaluated all approaches in two different settings. The *in-topic* scenario splits the data into training and test data, which leads to arguments of the same topic to appear in both training and test data. The *cross-topic* scenario aims at evaluating the generalization of the models, i.e., answering the question as to how good the performance of the models is on yet unseen topics and therefore is the more complex task. We further split the experiments in two-classes (Argument or NoArgument) and three-classes (PRO, CON, NON).

#	level	labels	sentences
1	sentence	NON	The opposition to uranium mining and nuclear power within Australia also has been linked with overseas activities .
2		PRO	The industry has shown that it can safely handle , transport and store the radioactive wastes generated by nuclear power .
3		CON	Increasing the amount of waste shipped , particularly in less secure countries , is seen as a significant increase in risk to nuclear terrorism .
4	token	NON	Not many countries have uranium mines and not all the countries have nuclear technology , so they have to hire both things overseas .
5		PRO	80 percent agreed that carbon - free nuclear energy should be expanded as one way to reduce greenhouse gases and prevent global climate change .
6		PRO & CON	Nuclear energy may have horrific consequences if an accident occurs , but it has an enormous capacity for energy production with no carbon emissions .

Fig. 2 Example sentences with annotations for the topic *"nuclear energy"* from sentence- [28] and token-level [31] datasets

Table 1 Sentence-level Macro-F_1 score for 2 classes (argumentative, non-argumentative) and for 3 classes (PRO, CON, NON) for the in-topic and cross-topic setups from our previous publication [9]

	Method	In-Topic	Cross-Topic
2-classes	BiLSTM	0.74	0.66
	BiCLSTM	0.74	0.70
	BiLSTM-KG	0.73	0.68
	CAM-Bert	0.80	0.67
	TACAM-Bert	**0.81**	**0.80**
3-classes	BiLSTM	0.56	0.46
	BiCLSTM	0.53	0.47
	BiLSTM-KG	0.56	0.47
	CAM-Bert	**0.73**	0.61
	TACAM-Bert	0.69	**0.64**

For all tasks we compare the following approaches:

- BiLSTM is the first baseline: a bidirectional LSTM model [13] that does not use topic information at all.
- BiCLSTM is the second baseline: a contextual bidirectional LSTM [10]. Topic information is used as an additional input to the gates of an LSTM cell. We use the version from [28] where the topic information is only used at the $i-$ and $c-$gates since this model showed the most promising results in their work.
- BiLSTM-KG is our bidirectional LSTM model using Knowledge Graph embeddings from DBPedia as the context source for the topic.
- CAM-Bert is our fine-tuning based transfer learning approach without topic information.
- TACAM-Bert is our fine-tuning based transfer learning approach with topic information.

Table 1 shows that for the in-topic scenario our models TACAM-Bert and CAM-Bert are able to improve the Macro-F_1 score by 7% for the two-class and by 17% for the three-class classification task by using context information from transfer learning compared to the previous state-of-the-art system BiCLSTM [28]. For the more complex cross-topic task we improve the two-class setup by 10% and for the three-class setup by 17%. Our experimental results show that considering topic and context information from pre-trained models improves upon state-of-the-art argument detection models considerably. The number of parameters of the models and the hyper parameters of the training are reported in the previous publication [9].

3.1.2 Token-Level Models

Our motivation for token-level, i.e., fine-grained, models is that they support more specific selection of argumentative spans within sentences. In addition, the shorter segments are better suited to be extracted and displayed in applications (e.g., argument search engines), which usually present arguments without surrounding context sentences.

We created a new token-level (fine-grained) corpus [31]. Crowdworkers had the task of selecting argumentative spans for a given set of topics and topic related sentences. The sentences were from textual data extracted from Common Crawl[6] for a predefined list of eight topics. The final annotations of five crowdworkers per sentence were merged and a label from the set {PRO, CON, NON} was assigned to each token (word) in the sentence. The final corpus, the AURC (argument unit recognition and classification) corpus, contains 8000 sentences with 4500 being argumentative sentences and a total of 4973 argumentative segments. Examples for token-level annotations of argumentative spans in the AURC corpus are displayed in Fig. 2 in lines 4–6.

The differentiator to previous work and datasets is that there are many sentences in AURC with more than one argumentative segment. An example for a sentence with mixed stance segments can be seen in Fig. 2 in line 6, with a CON and a PRO segment. This kind of fine-grained argumentative data cannot be modeled correctly with a sentence-level approach.

After the corpus creation process, we applied state-of-the-art models in natural language processing to establish strong baselines for this new task of AURC. The proposed baselines were a majority baseline (where all tokens were labeled with the most frequent class), a BiLSTM model (using the FLAIR library [1]) and a BERT model [7] in several configurations (such as base, large and with a CRF-layer). The performance of the models was compared with two different data splits. (i) An in-domain split, where the models were trained, evaluated and tested on the same set of topics. (ii) A cross-domain split, where the models were trained on a subset of the available topics and evaluated and tested on different out-of-domain topics. The second set-up is more challenging, since the models have to generalize the argument span selection for unseen topics. Furthermore, the cross-domain split is also closer to a real world application,

[6] http://commoncrawl.org/2016/02/february-2016-crawl-archive-now-available/.

Table 2 Token-level Macro-F_1 for 2 classes (2-cl: ARG, NON) and for 3 classes (3-cl: PRO, CON, NON) for the in-domain and cross-domain setups from our previous publication [31]

	Set	In-Domain	Cross-Domain
2-classes	dev	0.813	0.797
	test	0.782	0.770
3-classes	dev	0.743	0.615
	test	0.696	0.620

since we typically encounter topics that are not covered in the training set in many practical applications.

An interesting insight from this experiment is that it is also quite challenging for humans to correctly classify argumentative spans. It is probably for this reason that, depending on the evaluation measure, some models performed better than the human annotators. An error analysis provided the following interesting insights: The most common error was incorrect stance classification (especially in the cross-domain setup) compared to good performance for span recognition, for both in-domain and cross-domain. Table 2 shows the results for the best models.

In summary, token-level (i.e., fine-grained) models are close to or better than human performance for known topics. While the cross-domain setup turned out to be challenging, the results for in-domain topics are already useful and can be helpful for many downstream tasks in computational argumentation. Examples include clustering or grouping of similar arguments for the ranking task in argument search engines; and the summarization of argument segments in automated debating systems[7] that generate fluent compositions of extracted argumentative segments. Future work should address annotating sentences for many more topics, cross-domain performance and better representations for linguistic objects of different granularities.

3.2 Same-Side Stance Classification

As the experiments in our previous work ([9], see also Table 1) showed, there is still a huge gap of 16% Macro-F_1 score between the two-class and the three-class cross-topic scenario and of 8% in the in-topic scenario. The reason is that stance detection is a complex task. The Same-Side Stance Classification (SSSC) Challenge[8] addresses this problem. As an illustration consider the PRO argument "religion gives purpose to life". The PRO argument "religion gives moral guidance" is an example for a same-side argument, whereas the CON argument "religion makes people fanatic" is an example for a not-same-side argument.

Given two arguments regarding a certain topic, the SSSC task is to decide whether or not the two arguments have the same stance. This can be exploited for stance classification since the relations bring to bear additional information (information about the network of all arguments) for improved stance classification.

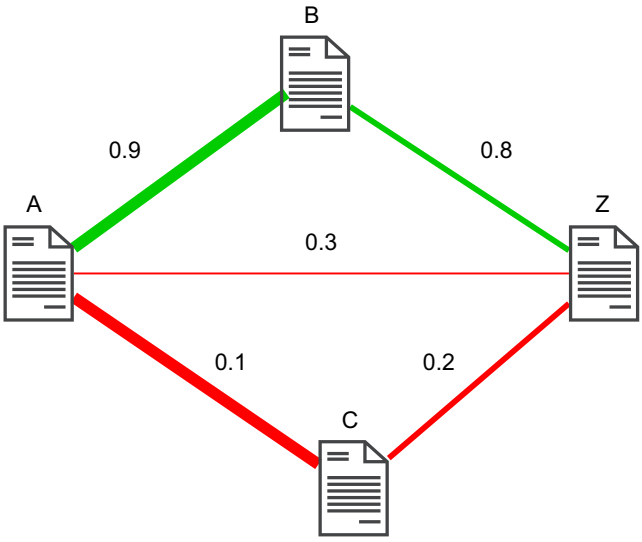

Fig. 3 Example of an argument graph. The nodes are represented as arguments and the edges as the binary SSSC relation. The thickness and the color of the edges represent the confidence and the class. Low confidence values can be interpreted as high confidence values against the relation

Our group participated in the challenge with a pretrained transformer model [7] fine-tuned on the SSSC data. We organized the data as graphs in the following way: we generated one graph per topic where the nodes are arguments and the edges are weighted with the confidence that the SSSC relation holds. If it is already known (e.g., from the training set) that the arguments agree or disagree, the confidence is 0 and 1 accordingly. Otherwise we use the probability predicted by the fine-tuned transformer model. Fig. 1 shows an illustration of the graph.

For each pair of arguments in the test set we computed the confidence of all paths of length k, and greedily selected the edge with the highest confidence for either an agreement or a disagreement between the two arguments. We computed the path score as the product of confidences of the edges on a path. By using the graph structure and the transitivity of the SSSC relation we could improve our Macro-F_1 score from 0.57 by 7 points for the cross-topic scenario.

4 Conclusion

Our ongoing work addresses several of the issues discussed in Sect. 3. Important issues we are addressing are the improvement of stance classification and the annotation for a larger number of topics. For stance classification, it is of

[7] https://www.research.ibm.com/artificial-intelligence/project-debater/.

[8] https://sameside.webis.de/.

interest to incorporate additional information in a multi-task learning setup, e.g., sentiment information and information from knowledge graphs. For annotating more topics, we can use our current models, which are trained on the eight AURC topics with gold labels, for a better sampling of sentences from a corpus such as OpenWebText [11] for new topics.

5 Future Work

This project overview mostly addressed lower-level tasks in computational argumentation. These are very important and essential to solve higher-level tasks that can only be accomplished with this extracted argumentative information on the sentence- and token-level. For the future we see these tasks as building blocks for high-level argumentation applications. One such application is argument validation, i.e., the classification of a sequence of two sentences as a valid vs. invalid link in a reasoning chain. With our improved argument mining techniques and based on our relational framework for stance classification, we would like to exploit graphs for argument validation. Another high-level argumentation application is interpretability of argument mining decisions: users in many applications can benefit from being able to view the rationale for why a particular sentence was selected as argumentative and with a particular stance. Here the human-interpretable information sources that we incorporated into sentence-level mining could be the basis for more effective methods. For future work, we are also considering other demanding tasks which could benefit from our work. One is the clustering or grouping of argumentative sentences or segments; and a second one the summarization of argument segments in automated debating systems that generate fluent compositions of extracted argumentative segments.

Acknowledgements This work has been funded by the Deutsche Forschungsgemeinschaft (DFG) within the project*ReMLAV*, Grant Numbers SCHU 2246/13 and SE 1039/10, as part of the Priority Program *RATIO*(SPP-1999). We are grateful to our collaborators Johannes Daxenberger, Christian Stab, Evgeniy Faerman and Iryna Gurevych.

Funding Open Access funding provided by Projekt DEAL.

References

1. Akbik A, Bergmann T, Blythe D, Rasul K, Schweter S, Vollgraf R (2019) Flair: an easy-to-use framework for state-of-the-art nlp. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics (Demonstrations)
2. Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, Ives Z (2007) DBpedia: a nucleus for a web of open data. In The semantic web. Springer, Berlin, Heidelberg, pp 722–735
3. Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. Trans Assoc Comput Linguist 5:135–146
4. Bollacker K, Evans C, Paritosh P, Sturge T, Taylor J (2008) Freebase: a collaboratively created graph database for structuring human knowledge. In: Proc. of the 2008 ACM SIGMOD Int. Conf. on Management of data, pp 1247–1250
5. Dai AM, Le QV (2015) Semi-supervised sequence learning. In Advances in neural information processing systems, pp 3079–3087
6. Daxenberger J, Eger S, Habernal I, Stab C, Gurevych I (2017) What is the essence of a claim? cross-domain claim identification. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp 2055–2066
7. Devlin J, Chang M, Lee K, Toutanova K (2018) BERT: pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol 1 (Long and Short Papers), pp 4171–4186
8. Freeman JB (2011) Argument structure: representation and theory, Vol 18. Springer Science & Business Media, Dordrecht
9. Fromm M, Faerman E, Seidl T (2019) TACAM: topic and context aware argument mining. In: 2019 IEEE/WIC/ACM Int. Conf. on Web Intelligence WI 2019, Thessaloniki, Greece, October 14–17, 2019, pp 99–106
10. Ghosh S, Vinyals O, Strope B, Roy S, Dean T, Heck LP (2016) Contextual LSTM (CLSTM) models for large scale NLP tasks. CoRR:(abs/1602.06291)
11. Gokaslan A, Cohen V (2019) Openwebtext corpus. http://web.archive.org/web/*/https://skylion007.github.io/OpenWebTextCorpus/
12. Habernal I, Gurevych I (2016) Argumentation mining in user-generated web discourse. Comput Linguist 43(1):125–179
13. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
14. Hua X, Wang L (2017) Understanding and detecting supporting arguments of diverse types. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Vol 2: Short Papers), pp 203–208
15. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint. arXiv:1301.3781
16. Nguyen H, Litman D (2015) Extracting argument and domain words for identifying argument components in texts. In: Proceedings of the 2nd Workshop on Argumentation Mining. Association for Computational Linguistics, Denver, CO, pp 22–28
17. Nguyen H, Litman D (2016) Context-aware argumentative relation mining. In: Long papers, vol 1. Association for Computational Linguistics, Berlin, pp 1127–1137
18. Nickel M, Murphy K, Tresp V, Gabrilovich E (2015) A review of relational machine learning for knowledge graphs. Proc IEEE 104(1):11–33
19. Nickel M, Tresp V, Kriegel H (2011) A three-way model for collective learning on multi-relational data. In: Proc. of the 28th Int.

Conf. on Machine Learning ICML 2011, Bellevue, Washington, USA, June 28–July 2, 2011, pp 809–816
20. Palau RM, Moens MF (2009) Argumentation mining: the detection, classification and structure of arguments in text. In: Proc. of the 12th Int. Conf. on Artificial Intelligence and Law ICAIL '09. ACM, New York, NY, USA, pp 98–107
21. Pennington J, Socher R, Manning C (2014) Glove: global vectors for word representation. In: Proc. of the 2014 Conf. on empirical methods in natural language processing (EMNLP), pp 1532–1543
22. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. In: Proc. of NAACL
23. Radford A, Narasimhan K, Salimans T, Sutskever I (2018) Improving language understanding by generative pre-training. https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/languageunderstandingpaper.pdf
24. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I (2019) Language models are unsupervised multitask learners. OpenAI Blog 1:8
25. Stab C, Daxenberger J, Stahlhut C, Miller T, Schiller B, Tauchmann C, Eger S, Gurevych I (2018) ArgumenText: Searching for arguments in heterogeneous sources. In: Proceedings of the 2018 Conf. of the North American Chapter of the Association for Computational Linguistics: Demonstrations. Association for Computational Linguistics, New Orleans, pp 21–25
26. Stab C, Gurevych I (2014) Identifying argumentative discourse structures in persuasive essays. In: Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing EMNLP 2014, Doha, Qatar, October 25–29, 2014, pp 46–56 (A meeting of SIGDAT, a Special Interest Group of the ACL)
27. Stab C, Gurevych I (2017) Parsing argumentation structures in persuasive essays. Comput Linguist 43(3):619–659
28. Stab C, Miller T, Gurevych I (2018) Cross-topic argument mining from heterogeneous sources using attention-based neural networks. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp 3664–3674
29. Suchanek FM, Kasneci G, Weikum G (2007) Yago: a core of semantic knowledge. In: Proc. of the 16th Int. Conf. on World Wide Web, pp 697–706
30. Toulmin SE (1958) The uses of argument. Cambridge University Press
31. Trautmann D, Daxenberger J, Stab C, Schütze H, Gurevych I (2020) Fine-grained argument unit recognition and classification. In: The Thirty-Fourth AAAI Conf. on Artificial Intelligence AAAI 2020, New York City, NY, USA. AAAI Press. https://aaai.org/Papers/AAAI/2020GB/AAAI-TrautmannD.7498.pdf
32. Wachsmuth H, Potthast M, Al Khatib K, Ajjour Y, Puschmann J, Qu J, Dorsch J, Morari V, Bevendorff J, Stein B (2017) Building an argument search engine for the web. In: Proceedings of the 4th Workshop on Argument Mining. Association for Computational Linguistics, Copenhagen, pp 49–59
33. Walton D (2012) Argument mining by applying argumentation schemes. Stud Log 4(1):2011

5 Diversity Aware Relevance Learning for Argument Search

The chapter includes the following publication:

> **Michael Fromm***, Max Berrendorf*, Sandra Obermeier, Thomas Seidl, and Evgeniy Faerman. „Diversity Aware Relevance Learning for Argument Search.“ In: *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part II*. vol. 12657. Lecture Notes in Computer Science. * equal contribution. Springer, 2021, pp. 264–271. DOI: 10.1007/978-3-030-72240-1_24. URL: https://doi.org/10.1007/978-3-030-72240-1_24

and the code is available at:

> https://github.com/fromm-m/ecir2021-am-search

Declaration of Authorship The research idea was proposed by Michael Fromm, developed and conceptualized by Michael Fromm and other co-authors. Michael Fromm, Max Berrendorf, and Sandra Obermeier did the implementation and the design of the architecture, the framework and the pipeline. Michael Fromm and Max Berrendorf conducted the experiments and analyzed their results. The findings were discussed with all authors. Michael Fromm, Evgeniy Faerman, Max Berrendorf, and Sandra Obermeier wrote the manuscript. All authors revised the manuscript.

Diversity Aware Relevance Learning for Argument Search

Michael Fromm*[1], Max Berrendorf*[1], Sandra Obermeier[1], Thomas Seidl[1], and Evgeniy Faerman[1]

Database Systems and Data Mining, LMU Munich, Germany
fromm@dbs.ifi.lmu.de

Abstract. In this work, we focus on retrieving relevant arguments for a query claim covering diverse aspects. State-of-the-art methods rely on explicit mappings between claims and premises and thus cannot utilize extensive available collections of premises without laborious and costly manual annotation. Their diversity approach relies on removing duplicates via clustering, which does not directly ensure that the selected premises cover all aspects. This work introduces a new multi-step approach for the argument retrieval problem. Rather than relying on ground-truth assignments, our approach employs a machine learning model to capture semantic relationships between arguments. Beyond that, it aims to cover diverse facets of the query instead of explicitly identifying duplicates. Our empirical evaluation demonstrates that our approach leads to a significant improvement in the argument retrieval task, even though it requires fewer data than prior methods. Our code is available at https://github.com/fromm-m/ecir2021-am-search.

Keywords: Argument Similarity · Argument Clustering · Argument Retrieval

1 Introduction

Argumentation is a paramount process in society, and debating on socially relevant topics requires high-quality and relevant arguments. In this work, we deal with the problem of *argument search*, which is also known as *argument retrieval*. The goal is to develop an Argument Retrieval System (ARS) which organizes arguments, previously extracted from various sources [4, 8, 15, 17], in an accessible form. Users then formulate a query to access relevant arguments retrieved by the ARS. The query can be defined as a *topic*, e.g. *Energy* in which case the ARS retrieves all possible arguments without further specification [10, 15, 17]. Our work deals with a more advanced case, where a query is formulated in the form of a *claim*, and the user expects *premises* attacking or supporting this query claim. An example of a claim related to the topic *Energy* could be *"We should abandon Nuclear Energy"* and a supporting premise, e.g., *"Accidents caused by Nuclear Energy have longstanding negative impacts"*. A popular search methodology to

* equal contribution

find relevant premises is a similarity search, where the representations of the retrieved premises are similar to the representation of the (augmented) query claim [1, 3, 9, 16]. However, as noted by [6, 7], the relevance of a premise does not necessarily coincide with pure text similarity. Therefore, the authors of [6] advocate to utilize the similarity between the query claim and other claims in an ARS database and retrieve the premises assigned to the most similar claims. However, such ARS requires ground truth information about the premise to claim assignments and therefore has limited applicability: Either the information sources are restricted to those sources where such information is already available or can automatically be inferred, or expensive human annotations are required. To mitigate this problem and keep the original system's advantages, we propose to use a machine learning model *to learn* the relevance between premises and claims. Using this model, we can omit the (noisy) claim-claim matching step and evaluate the importance of (preselected) candidate premises directly for the query claim. Since the relevance is defined on the semantic level, we have to design an appropriate training task to enable the model to learn semantic differences between relevant and non-relevant premises. Furthermore, an essential subtask for an ARS is to ensure that the retrieved premises do not repeat the same ideas. Previous approaches [6] employ clustering to eliminate duplicates. However, clustering approaches often group data instances by other criteria than expected by the users [12], as also observed in Argument Mining (AM) applications [13]. For our method, we propose an alternative to clustering based on the idea of *core-sets* [14], where the goal is to cover the space of relevant premises as well as possible.

2 Preliminaries

In our setting, the query comes in the form of a claim, and an answer is a sorted list of *relevant* premises from the ARS database. A premise is considered relevant if it attacks or supports the idea expressed in the claim [11, 19]. We denote the query claim by c_{query} and the list of premises retrieved by ARS by A, with the length being fixed to $|A| = k$. Besides relevance, another vital requirement for the ARS is that premises in A should have diverse semantic meaning. We consider a two-step retrieval process. First, in the *pre-filtering*, the system selects a set of candidate premises $\mathcal{T}$ with $|\mathcal{T}| > k$. This step should have a relatively high recall, i.e., find most of the relevant premises. For a fair comparison to previous approaches, we leave the pre-filtering step from [6] unchanged. We note that the current version of pre-filtering requires ground-truth matchings of premises to claims restricting its applicability and improving it in future work. The pre-filtering process described in [6] has several steps. When a query claim arrives, the system first determines *claims* from the database which have the highest Divergence from Randomness [2] similarity to the query claim. Next, the system receives the corresponding claim clusters of the claims found in the previous step, and all premises assigned to all claims from these clusters are collected in a candidate seed set $\mathcal{T}_{seed}$. Each premise $p \in \mathcal{T}_{seed}$ is then used as a query to obtain

the most similar premises using the BM25 score, which are accumulated in a set $\mathcal{T}_{sim}$. The complete candidate set is then given as the union $\mathcal{T} = \mathcal{T}_{seed} \cup \mathcal{T}_{sim}$.

3 Our Approach for Candidate Refinement

Our work's primary focus is the second step in the retrieval process or the candidate refinement/ranking procedure. The candidates are analyzed more thoroughly in the refinement step, and non-relevant or redundant premises are discarded. Our refinement process comprises two components. The *relevance filter* component determines each premise's relevance from the candidate set $\mathcal{T}$ using an advanced machine learning model that keeps only the most relevant ones. The relevance filter thus maps the candidate set $\mathcal{T}$ to a subset thereof, denoted by $\mathcal{T}_{filtered} \subseteq \mathcal{T}$. The subsequent *premise ranker* selects and orders k premises from $\mathcal{T}_{filtered}$ to the result list A. An essential requirement for the premise ranker is that A does not contain semantically redundant premises. In the following, we describe both components in more detail.

3.1 Relevance Filter

Inference Given a set of candidate premises $\mathcal{T}$ and the query claim c_{query}, the *relevance filter* determines the relevance score of each candidate $p \in \mathcal{T}$ denoted as $r(p \mid c_{query})$. We keep only the most relevant candidates in the filtered candidate set $\mathcal{T}_{filtered} = \{p \in \mathcal{T} \mid r(p \mid c_{query}) > \tau\}$ with a relevance threshold τ. We interpret the relevance prediction as a binary classification problem and train a Transformer [18] model to solve this classification task given the concatenation of the candidate premise and the query claim. At inference time, we use the predicted likelihood as the relevance score and evaluate the model on the concatenation of each candidate premise with the query claim.

Training Task For the training part, we assume that we have access to a (separate) dataset $D = (\mathcal{P}', \mathcal{C}', \mathcal{R}^+)$ containing a set of premises $\mathcal{P}'$, a set of claims $\mathcal{C}'$ and a set of relevant premise-claim pairs $\mathcal{R}^+ \subseteq \mathcal{P}' \times \mathcal{C}'$. In fact, several datasets fulfill this requirement, e.g., [7, 20]. Since the relevance filter receives as input the remaining candidate premises after the pre-filtering, we assume that the non-relevant premises appear similar to the relevant ones. Therefore, the training task must be designed very carefully to enable the model to learn semantic differences between relevant and non-relevant premises. We use the ground truth premise-claim pairs $\mathcal{R}^+$ as instances of the positive class (i.e., an instance of matching pairs). For each positive instance $(p^+, c) \in \mathcal{R}^+$, we generate L instances of the negative class $(p_i^-, c) \in \mathcal{R}^-$. For p_i^-, we choose the L most similar premises according to a premise similarity $psim$, which do not co-occur with c in the database. We use the cosine similarity $psim(p, p') = \cos(\phi(p), \phi(p'))$ between the premise representations $\phi(p)$ obtained from a pre-trained BERT model without any fine-tuning as premise similarity.[1] The transformer model,

[1] Using average pooling of the second-to-last hidden layer over all tokens

Algorithm 1: Biased Coreset

Data: candidates $\mathcal{T}$, relevances R, similarity $psim$, $k \in \mathbb{N}$, $\alpha \in [0, 1]$
Result: premise list A
for $i = 1$ **to** k **do**
  if $|A| = 0$ **then** $a = \underset{p \in \mathcal{T}}{\operatorname{argmax}}\, \alpha \cdot R[p]$;
  else $a = \underset{p \in \mathcal{T}}{\operatorname{argmax}}\, \alpha \cdot R[p] - (1 - \alpha) \cdot \max_{a \in A} psim(a, p)$;
  $A.append(a)$; $\mathcal{T} = \mathcal{T} \setminus \{a\}$
end

which predicts the premise-claim relevance, is initialized with weights from a pre-trained BERT model [5].

3.2 Premise Ranker

The *premise ranker* receives a set of relevant premises with the corresponding relevance scores and makes the final decision about the premises and the order they are returned to the user. Since the two relevance filtering steps have been applied, we assume that most remaining candidates are relevant. Thus, the main task of this component is to avoid semantic duplicates. While related approaches [6] advocate for the utilization of clustering for the detection of duplicates and expect that premises with the same meaning end up in the same clusters, we pursue a different idea. Instead of explicitly detecting the duplicates, we aim to identify k premises that adequately represent all premises in $T_{filtered}$. Therefore, we borrow the idea of core-sets from [14] and aim to select k premises from the final candidate set $\mathcal{T}_{filtered}$ such that for each candidate premise $p \in \mathcal{T}_{filtered}$ there is a similar premise in the result A. More formally, we denote $Q(p, A) = \max_{a \in A} psim(p, a)$ as a measure of how well p is represented by A, using the premise similarity $psim$. Thus, $\bar{Q}(A) = \min_{p \in \mathcal{T}_{filtered}} Q(p, A)$ denotes the worst representation of any premise $p \in \mathcal{T}_{filtered}$ by A. Hence, we aim to maximize $\bar{Q}$ such that every premise p is well represented. This min-max objective ensures that every premise is well-represented at not only the majority of premises. To solve the selection problem, we adopt the greedy approach from [14]. Since our goal is not only that the selected premises represent the remaining candidates well, but also that the selected premises have high relevance, we start with the most relevant premise and also consider the relevance score r for the next assignments, with a weighting parameter $\alpha \in [0, 1]$. $\alpha = 0$ scores only according to the coreset criterion, while $\alpha = 1$ uses only the relevance. The full algorithm is presented in Algorithm 1.

Premise Representation The premise ranker requires a meaningful similarity measure to compare premises with each other. As also noted in [6], semantically similar premises might often be expressed differently. Therefore, an essential requirement for the similarity function is that it captures semantic similarities.

We investigate two approaches to obtain vector representations on which we compute similarities using l1, l2, or cos similarity. Previous works demonstrated that BERT models pre-trained on language modeling can capture argumentative context [10]. Thus, our first *BERT* similarity function employs a BERT model without fine-tuning to encode the premises. We abbreviate these representations with *BERT*. As an alternative, we propose representing each premise by a vector of relevance scores to selected claims in the database. While we can use randomly selected claims or cluster all claims in the database, many databases already contain topic information about the claims, such as e.g., "Energy." Thus, we restrict the selection of claims for each premise to the same high-level topic of interest. In this case, all premises retrieved for a single query belong to the same topic. We do not consider it a substantial restriction since arguments always exist in some context, and it rarely makes sense to retrieve premises from different topics for the same query. We utilize our relevance filter model to compute relevance scores for the premise and each of the selected claims. We call the resulting vector of stacked similarities *CLAIM-SIM* representation. We hypothesize that a similar relationship to the selected claims is a good indicator of semantically similar premises.

4 Evaluation

Experimental Setting The training dataset of the relevance filter is a subset of 160,000 positive (relevant) claim-premise sentence pairs of the dataset described in [7]. Additionally, we generated 320,000 negatives (not-relevant) claim-premise pairs as described in Section 3.1. For the evaluation of our approach and comparison with the baselines, we utilize the dataset from [6]. The evaluation set consists of 1,195 triples $(c_{query}, c_{result}, p_{result})$ each labeled as "very relevant" (389), "relevant" (139) or "not relevant" (667). The 528 "very relevant" and "relevant" premises were assigned to groups with the same meaning by human annotators. In contrast to [6] we do *not* utilize the ground truth assignments of $c_{result} \leftrightarrow p_{result}$ in our approach. Therefore our method can utilize newly arriving premises without an assignment to c_{result}. To select the optimal hyperparameters for our approach and avoid test leakage, we use leave-one-out cross-validation: For each query claim with corresponding premises, we use the rest of the evaluation dataset to select the hyperparameters and then evaluate this hold-out query. To obtain a final score, we average over all splits. As an evaluation metric, we use the modified nDCG from [6]: Only the first occurrence from a premise ground truth cluster yields positive gain; duplicates do not give

Table 1. Modified NDCG score for $k = 5$ and $k = 10$.

	[6]			top-k			k-Means		Biased Coreset	
k	first	sent	sliding	zero-shot	same topic	ours	BERT	CLAIM-SIM	BERT	CLAIM-SIM
5	.399	.378	.455	.437	.373	.447	.428	.465	.437	**.475**
10	.455	.429	.487	.476	.448	.502	.515	.513	.520	**.526**

any gain. In Table 1, we summarize the results of the argument retrieval task. The numbers represent the modified NDCG scores for $k = 5$ and $k = 10$. The first three columns show the evaluation results for the methods from [6].[2] In the next three columns denoted as *top-k*, we present the results when premises with the highest score are returned directly, without de-duplication. With the *zero-shot* approach, we investigate the assumption that similarity between query and claim is not a sufficient indicator for relevance. Thus, we use the similarity between representations obtained from a pre-trained BERT model without training on claim-premise relevance. The second column, *same topic*, denotes the performance of the relevance model trained in the same setting as our approach with the only difference that negative instances for the training are selected from the same topic. Finally, *ours* denotes the setting, where k instances have the highest probability to be relevant estimated by our model (more precisely, the *relevance filter*). Given these results, we observe a strong performance of the *zero-shot* approach, which comes close to the approaches by [6]. We emphasize that this is even though this baseline approach neither uses ground truth premise-claim relevance data as [6], nor any other external premise-claim relevance data. Moreover, we observe that we can achieve good performance in terms of the *modified* NDCG despite not filtering duplicates. At the same time, we observe that our model can still improve the similarity-based approach by several points. In contrast, the model learned with negatives instances from the same topic performs much worse than *zero-shot*, which underlines the correct task's importance. Finally, the columns denoted as *Biased Coreset* present our final results. The results are from the *premise ranker* applied to the different premise representations of the most relevant premises selected by *relevance filter*. For comparison, we also report the results, where k-means is used as *premise ranker* on the same representations, where we select at most one premise per cluster according to the similarity. The *claim-sim* premise representation always outperforms *bert* and our *biased-coreset* premise ranker is better than the k-means clustering.

5 Conclusion

In this work, we have presented a novel approach for the retrieval of *relevant* and *original* premises for the query claims. Our new approach can be applied more flexibly than previous methods since it does not require mappings between premises and claims in the database. Thus, it can also be applied in an inductive setting, where new premises can be used without the need first to associate them with relevant claims manually. At the same time, it achieves better results than approaches that make use of this information.

[2] For the evaluation, we have used interim results provided by the authors of the original publication. Since we had obtained deviations from the originally reported results, we have contacted the authors and came together to the conclusion that our numbers are correct. We thank the authors for their help.

6 Acknowledgments

This work has been funded by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A and by the Deutsche Forschungsgemeinschaft (DFG) within the project Relational Machine Learning for Argument Validation (ReMLAV), Grant NumberSE 1039/10-1, as part of the Priority Program "Robust Argumentation Machines (RATIO)" (SPP-1999). The authors of this work take full responsibility for its content.

References

1. Akiki, C., Potthast, M.: Exploring Argument Retrieval with Transformers. In: Working Notes Papers of the CLEF 2020 Evaluation Labs (Sep 2020)
2. Amati, G., Van Rijsbergen, C.J.: Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Transactions on Information Systems (TOIS) **20**(4), 357–389 (2002)
3. Bondarenko, A., Fröbe, M., Beloucif, M., Gienapp, L., Ajjour, Y., Panchenko, A., Biemann, C., Stein, B., Wachsmuth, H., Potthast, M., Hagen, M.: Overview of Touché 2020: Argument Retrieval. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) Working Notes Papers of the CLEF 2020 Evaluation Labs. CEUR Workshop Proceedings, vol. 2696 (Sep 2020), http://ceur-ws.org/Vol-2696/
4. Chernodub, A., Oliynyk, O., Heidenreich, P., Bondarenko, A., Hagen, M., Biemann, C., Panchenko, A.: Targer: Neural argument mining at your fingertips. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 195–200 (2019)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://www.aclweb.org/anthology/N19-1423
6. Dumani, L., Neumann, P.J., Schenkel, R.: A framework for argument retrieval. In: European Conference on Information Retrieval. pp. 431–445. Springer (2020)
7. Dumani, L., Schenkel, R.: A systematic comparison of methods for finding good premises for claims (2019)
8. Ein-Dor, L., Shnarch, E., Dankin, L., Halfon, A., Sznajder, B., Gera, A., Alzate, C., Gleize, M., Choshen, L., Hou, Y., et al.: Corpus wide argument mining-a working solution. In: AAAI. pp. 7683–7691 (2020)
9. Feger, M., Steimann, J., Meter, C.: Structure or content? towards assessing argument relevance. In: Proceedings of the 8th International Conference on Computational Models of Argument (COMMA 2020). p. 135 (2020)
10. Fromm, M., Faerman, E., Seidl, T.: TACAM: topic and context aware argument mining. In: Barnaghi, P.M., Gottlob, G., Manolopoulos, Y., Tzouramanis, T., Vakali, A. (eds.) 2019 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2019, Thessaloniki, Greece, October 14-17, 2019. pp. 99–106. ACM (2019). https://doi.org/10.1145/3350546.3352506, https://doi.org/10.1145/3350546.3352506

11. Habernal, I., Gurevych, I.: Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1589–1599. Association for Computational Linguistics, Berlin, Germany (Aug 2016). https://doi.org/10.18653/v1/P16-1150, https://www.aclweb.org/anthology/P16-1150
12. Kriegel, H.P., Kröger, P., Zimek, A.: Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. ACM Transactions on Knowledge Discovery from Data (TKDD) **3**(1), 1–58 (2009)
13. Reimers, N., Schiller, B., Beck, T., Daxenberger, J., Stab, C., Gurevych, I.: Classification and clustering of arguments with contextualized word embeddings. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 567–578. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1054, https://www.aclweb.org/anthology/P19-1054
14. Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. arXiv preprint arXiv:1708.00489 (2017)
15. Stab, C., Miller, T., Schiller, B., Rai, P., Gurevych, I.: Cross-topic argument mining from heterogeneous sources. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 3664–3674. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018). https://doi.org/10.18653/v1/D18-1402, https://www.aclweb.org/anthology/D18-1402
16. Staudte, C., Lange, L.: Sentarg: A hybrid doc2vec/dph model with sentiment analysis refinement. In: CLEF (2020)
17. Trautmann, D., Fromm, M., Tresp, V., Seidl, T., Schütze, H.: Relational and fine-grained argument mining. Datenbank-Spektrum pp. 1–7 (2020)
18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
19. Wachsmuth, H., Naderi, N., Hou, Y., Bilu, Y., Prabhakaran, V., Thijm, T.A., Hirst, G., Stein, B.: 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. pp. 176–187 (2017)
20. Wachsmuth, H., Potthast, M., Al Khatib, K., Ajjour, Y., Puschmann, J., Qu, J., Dorsch, J., Morari, V., Bevendorff, J., Stein, B.: Building an argument search engine for the web. In: Proceedings of the 4th Workshop on Argument Mining. pp. 49–59 (2017)

6 Active Learning for Argument Strength Estimation

The chapter includes following publication:

> Nataliia Kees, **Michael Fromm**, Evgeniy Faerman, and Thomas Seidl. „Active Learning for Argument Strength Estimation.“ In: *Proceedings of the Second Workshop on Insights from Negative Results in NLP*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 144–150. URL: https://aclanthology.org/2021.insights-1.20

and the code is available at

> https://github.com/fromm-m/active-learning-argument-strength

Declaration of Authorship The research idea was proposed by Michael Fromm, developed and conceptualized by Michael Fromm and Evgeniy Faerman. Nataliia Kees did the implementation of the architecture. Michael Fromm, Nataliia Kees and Evgeniy Faerman designed and conducted the experiments. All authors discussed the findings. Michael Fromm and Nataliia Kees wrote the manuscript.

Active Learning for Argument Strength Estimation

Nataliia Kees and **Michael Fromm** and **Evgeniy Faerman** and **Thomas Seidl**
LMU Munich, Germany
kees.nataliia@gmail.com, fromm@dbs.ifi.lmu.de

Abstract

High-quality arguments are an essential part of decision-making. Automatically predicting the quality of an argument is a complex task that recently got much attention in argument mining. However, the annotation effort for this task is exceptionally high. Therefore, we test uncertainty-based active learning (AL) methods on two popular argument-strength data sets to estimate whether sample-efficient learning can be enabled. Our extensive empirical evaluation shows that uncertainty-based acquisition functions can not surpass the accuracy reached with the random acquisition on these data sets.

1 Introduction

Argumentative quality plays a significant role in different domains of social activity where information and idea exchange are essential, such as the public domain and the scientific world. Theoretical discussions about what constitutes a good argument can be traced back to the ancient Greeks (Smith, 2020). Researchers nowadays continue exploring this topic, trying out approaches that employ empirical machine learning estimation techniques (Simpson and Gurevych, 2018).

One of the most expensive and time-consuming tasks for machine learning-driven argument strength prediction is data labeling. Here, the result is highly dependent on the quality of labels, while the annotation task demands cognitive and reasoning abilities. One way to guarantee good annotations is to perform labeling with schooled experts, raising project costs extensively. For this reason, a common approach involves employing crowd workers. As argument strength detection is a highly subjective task, crowd workers' labeling results are often identified by low reliability and prompt researchers to counter-check the results with more crowd workers and as specifically developed agreement-based techniques. Sometimes a threshold for agreement cannot be reached at all, which might lead to data loss (see e.g. (Habernal and Gurevych, 2016a; Toledo et al., 2019).

This motivates us to investigate the applicability of some existing methods for reducing the amount of training data for automatic argument strength prediction. To this end, we look closely at the technique of active learning (AL). In this paper, we evaluate standard uncertainty-based acquisition functions for the argument strength prediction. We perform several experiments for the task of binary argument-pair classification (see Table 1) with several uncertainty-based data selection rounds. Our findings show that uncertainty-based AL techniques do not provide any advantages compared to random selection strategies. The cold-start problem and unreliable nature of annotations concerning argument strength might constitute the reasons for the failure of these techniques.

Argument 1	**Argument 2**
School uniforms are a BAD idea. I'm to lazy to explain it but trust me, I wore them 4 years.	School uniform cant save person out of cold or heat like special clothes. It is not comfortable when you sit for an hours in a classroom.

Table 1: Example of an argument pair both arguing against school uniforms (Habernal and Gurevych, 2016b)

2 Related Work

2.1 Argument Quality Estimation

In general, there is no agreement on how to operationalize argumentation quality (Toledo et al., 2019; Wachsmuth et al., 2017; Simpson and Gurevych, 2018; Persing and Ng, 2015; Lauscher et al., 2020). In some studies, argument strength is regarded in

Proceedings of the Second Workshop on Insights from Negative Results in NLP, pages 144–150

ISBN 978-1-954085-93-0

its persuasiveness and quantified as the proportion of people persuaded by the given argument (Habernal and Gurevych, 2016b; Persing and Ng, 2015; Toledo et al., 2019). Persuasiveness makes argument strength easy to operationalize and serves as a way of dealing with the unclear nature of the concept by approximating its meaning through relying on the majority's wisdom. This approach lies at the center of the crowd-sourcing data labeling efforts and is the most common approach undertaken in existing data sets. This limits the reliability of the labels attained in such a manner, though, due to the highly subjective nature of such labels.

2.2 Active Learning

Active learning is defined as a machine learning technique designed to assist in annotating unlabelled data sets by automatically selecting the most informative examples, which are subsequently labeled by human experts (the so-called *oracles*) (Hu, 2011; Cohn et al., 1996). A popular approach to estimating the informativeness of single data points involves quantifying model uncertainty from a sample of stochastic forward passes for a given data point. Common techniques such as entropy, mutual information, or variation ratios (see Appendix A.1 for more details) reportedly help reach good results on a range of tasks on high-dimensional data, e.g., in Computer Vision or Natural Language Processing (Gal et al., 2017; Siddhant and Lipton, 2018; Hu, 2011). The assumption behind this is that in this way, data points which are closest to the decision boundary can be selected, helping to fine-tune the line dividing the classes most efficiently.

So far, AL in argument mining has received little attention. In the work of (Ein-Dor et al., 2020), the authors propose an Iterative Retrospective Learning (IRL) variant for the argument mining task. Their approach, however, is focused on solving the class imbalance problem between arguments and non-arguments and is *precision-* rather than *accuracy-oriented* as AL is. Another approach is suggested by (Simpson and Gurevych, 2018). They apply the Gaussian process preference learning (GPPL) method for performing AL for estimating argument convincingness, which the authors expect to be helpful against the cold-start problem.

3 Data Set

For our analysis, we use two publicly available data sets suitable for the task of pairwise argument strength prediction:

- *UKPConvArg1Strict*, published by (Habernal and Gurevych, 2016b), consists of 11,650 argument pairs distributed over 16 topics.
- *IBM-9.1kPairs*, presented by (Toledo et al., 2019) consists of 9,125 argument pairs distributed over 11 topics.

Because supporting and opposing arguments often share the same vocabulary and semantics, we do not treat each stance within a given topic as a separate topic, contrary to the authors of the two data sets. Instead, we combine the "for" and "against" arguments within the same topic under the same topic index and, thus, avoid leakage of semantic information between train and test data split. This preprocessing makes the performance of our models not directly comparable with the performance from the original papers. However, reproducibility of the original papers' results is beyond the scope of this work, as our focus lies on testing AL acquisition functions instead of reaching higher performance with our models.

Due to the high computational costs of the AL process, we decide to select the three most representative topics from each data set. One way to reach high representativeness would be to select topics that are average in difficulty. Since we try to approximate a real-world setting where the labels are unknown, it is not clear at the beginning which topics are more challenging to learn than the others. For this reason, we decide to select our test topics according to their size. Thus, we cross-validate our models on each data set's smallest topic, the largest one, and the median-sized one. Thus, the topics we select according to this procedure are topics 10 ("Is the school uniform a good or bad idea?"), 13 ("TV is better than books") and 14 ("Personal pursuit or advancing the common good?") in *UKPConvArg1Strict* and topics 3 ("Does social media bring more harm than good?"), 4 ("Should we adopt cryptocurrency?") and 7 ("Should we ban fossil fuels?") in *IBM-9.1kPairs* data.

4 Experimental Setting

4.1 Research Design

This study aims to test the hypothesis that *uncertainty-based data acquisition strategies can help to achieve a better model performance than a mere random selection of the data for argument*

strength estimation. We test this by comparing different data selection strategies against random data selection, serving as a baseline.

We test our acquisition strategies on a task of pairwise (relative) argument strength comparison, constructed as a binary classification task, for which we use the *UKPConvArg1Strict* (Habernal and Gurevych, 2016b) and *IBM-9.1kPairs* (Toledo et al., 2019) data sets. The code to our experiments is publicly available.[1]

In order to employ uncertainty-based acquisition functions, we need to measure model uncertainty at prediction time. This is possible either by using Bayesian methods or by approximating their effect via obtaining distributions for output predictions by some other means. Based on the ground work layed out by (Gal and Ghahramani, 2016), who show that dropout training in deep neural networks help approximate Bayesian inference in deep Gaussian processes, we design our experiments as MC dropout. With this, we simulate several stochastic forward passes through the model at prediction time and sample repeatedly from softmax outputs to obtain prediction distributions.

4.2 Method and Procedure

Similar to the procedure stipulated by (Toledo et al., 2019), we fine-tune the pre-trained BERT-Base Uncased English (Devlin et al., 2018) for the task of binary argument-pair classification by adding a single classification layer on top. The BERT architecture includes dropout layers with a probability of 0.1 (Devlin et al., 2018). We keep it this way, which allows us to approximate model uncertainty as described above and test the uncertainty-based acquisition functions on the fine-tuned BERT-based. To do that, we enable dropout at inference time.

In order to estimate topic difficulty and validate our topic selection procedure described above, we train and test the models on all available labels of both data sets separately with the method of k-fold cross-validation, where k stands for the respective number of topics in a given data set. We separate every topic and use it as test data, with model training performed on the rest of the data, which helps to isolate the topics and measure their respective difficulty.

Our active learning experiments are conducted in a setting of a 3-fold cross-validation, with 3 indicating the number of most representative topics selected by us from the given data sets, as mentioned in Section 3. Thus, in each fold in our experiments, we test on one of the three selected topics for each data set (holdout data) and train on the rest of the compete data set (train-dev).

The train-dev data in each fold consists of random splits into train (85%) and validation (15%) data, whereas the validation, or development, data are used for measuring the goodness of fit of the model trained on the training data. Having separated and fixed the validation data, a batch of 130 argument pairs is selected randomly from the train split. These data are used as initial training data on which bert-base-uncased is fine-tuned according to our classification task.

Model evaluation is performed via accuracy measurement. Training on each of the three folds per data set is conducted ten times for improved reliability of the results. Thus, for each fold, we produce ten validation splits and ten initial training data batches to add some randomness into the experiments but in a controlled manner. They are kept fixed for every training fold to control for the effect of random initial data selection and enable a reliable comparison between the acquisition functions.

We add another 130 argument pairs in each learning round and re-train the fine-tuned model. Within this setting, the whole data set would be selected within approx. 55 iterations for *IBM-9.1kPairs* data and approx. 72 iterations for *UKPConvArg1Strict* data (when calculated with the median-sized test split size). In an attempt to minimize the burden associated with heavy training, we decide to limit each active learning process to (less than) a half iterations, stopping at the 27$^{\text{th}}$ iteration.

Further details on the hyperparameters and the computing and software infrastructure can be found in Appendix sections A.2 and A.4.

4.3 Acquisition Functions

We perform AL on three uncertainty-based acquisition functions one by one. In particular, we compare the performance of variation ratios, entropy, and BALD (Houlsby et al., 2011; Gal et al., 2017) against a random acquisition baseline. For each of the learning rounds, we acquire data based on the heuristics calculated over a sample of 20 stochastic forward pass outputs. Our expectation is that other measures will outperform the random acquisition.

[1] https://github.com/nkees/active-learning-argument-strength

5 Results

For the estimation of the performance of the models trained on the whole data with k-fold cross-validation, we reach a comparable performance of our BERT-based binary classification technique on both of the data sets (average accuracy on *UKP-ConvArg1Strict*: 0.76, on *IBM-9.1kPairs*: 0.77). This is a slightly worse performance than (Toledo et al., 2019) achieves with the same architecture; the reason could be attributed to a different topic attribution strategy, as well as to some differences in the used hardware or hyperparameters, such as batch size or the number of epochs.

We find that the topics selected by us from the *UKPConvArg1Strict* stand rather on the low end of difficulty, with model accuracy tending towards the upper end of the scale when validated on these topics: all of them are higher than the mean performance of 0.76 (see Appendix A.3 for more details). However, from the distribution point of view, two of the topics, namely 10 and 13, yield median model performance, making them, in our opinion, suitable representatives of the whole data.

As for the *IBM-9.1kPairs* data set, our selected topics produce on average comparable performance with the model performance on the whole topic set (accuracy of 0.776 vs. 0.77 respectively). They also represent the most difficult topic, the easiest topic, and one closely neighboring the median topic (accuracy of 0.78 being slightly higher than the median performance of 0.77). In this case, the selected topics provide a better representation of the whole data set and grant strong validity when it comes to generalizing the results of our experiments.

The series of experiments we conducted in order to test whether our proposed heuristics for AL data acquisition provide us with any significant improvement surprisingly do not reveal any heuristic which would perform better than in the case of a random acquisition. This is true both for *UKP-ConvArg1Strict* and *IBM-9.1kPairs* data; a detailed overview is presented in Tables 2 and 3. Statistical significance of the results has been tested with a Wilcoxon signed-rank test, which provides a non-parametric alternative to the paired T-test and is more suitable due to the non-Gaussian distribution of the differences in the results.

All heuristics result in performance that is lower than that of the random baseline. All of our results are statistically significant with p-values ≤ 0.0001.

Despite the fact that random acquisition turns out to be the best one in terms of performance, with our results being consistent through both data sets and the difference being statistically significant, it is still noticeable that the differences in each case are rather small (see Figures 1 and 2 for graphic visualization of the model performance during active learning rounds comparing the acquisition functions).

Heuristic	Mean	Variat.	Avg.Diff.
random (b.)	0.747	0.0881	-
entropy	0.7388	0.0925	-0.0082
variation ratios	0.7368	0.0922	-0.0103
bald	0.7377	0.0928	-0.0093

Table 2: Results of active learning experiments on *UKPConvArg1Strict*. Abbreviations: *b.* stands for *baseline*, *variat.* stands for *variation*, *avg.diff.* stands for *average difference*. Negative average difference means that the challenger heuristic has not outperformed the baseline.

Heuristic	Mean	Variat.	Avg.Diff.
random (b.)	0.7491	0.0855	-
entropy	0.7414	0.0878	-0.0077
variation ratios	0.7377	0.0923	-0.0114
bald	0.7412	0.0882	-0.0079

Table 3: Results of active learning experiments on *IBM-9.1kPairs*. Abbreviations: *b.* stands for *baseline*, *variat.* stands for *variation*, *avg.diff.* stands for *average difference*. Negative average difference means that the challenger heuristic has not outperformed the baseline.

6 Discussion

The results of our experiments do not point to any acquisition functions which outperform random acquisition. This finding does not exclude the possible existence of some other suitable acquisition functions, even from the same class (such as uncertainty-based). This remains an open question and should be considered in further research on the topic. For the time being, the random acquisition should be considered the approach of choice when selecting data for labeling for the task of pairwise argument strength prediction. This is sensible both from an accuracy standpoint as well as due to the computational cheapness of a random process.

As the literature suggests, a possible reason why uncertainty-based methods perform so unimpressively is their proneness to picking outliers – a disadvantage that some other methods, such

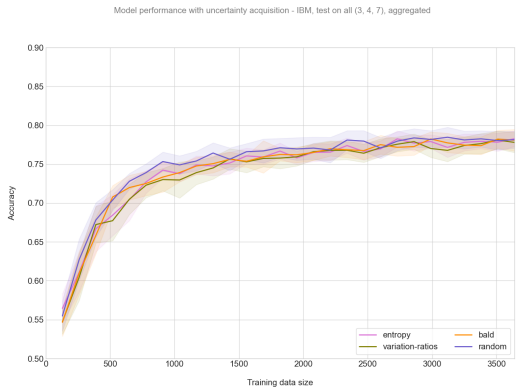

Figure 1: Overview of the training results on the *UKP-ConvArg1Strict* dataset based on different uncertainty-based acquisation methods

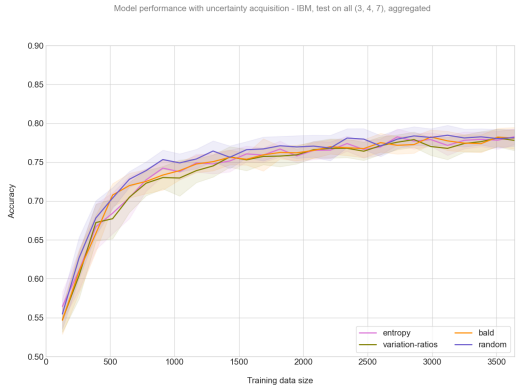

Figure 2: Overview of the training results on the *IBM-9.1kPairs* dataset based on different uncertainty-based acquisition methods

as diversity-based acquisition (e.g., (Sener and Savarese, 2018), do not have. This might be especially critical in the realm of argument strength prediction, as outliers might represent the arguments where relative argument strength difference is marginal, the data are noisy, or where the provided labeling is too subjective. Another critical factor is the cold-start problem, i.e., overfitting on the small initial data set of data, for which no initial informativeness estimation could be performed. This poses a drawback for the uncertainty-based methods, relying on the initial data sample for subsequent data acquisition.

7 Conclusion

This paper evaluates the effect of uncertainty-based acquisition functions, such as variation ratios, entropy, and BALD, on the model performance in the realm of argument strength prediction. As no acquisition function tested helps improve model performance in comparison to the random acquisition, we have not found any justification for using uncertainty-based active learning for pairwise argument strength estimation.

8 Acknowledgments

This work has been funded by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A and by the Deutsche Forschungsgemeinschaft (DFG) within the project Relational Machine Learning for Argument Validation (ReMLAV), Grant Number SE 1039/10-1, as part of the Priority Program "Robust Argumentation Machines (RATIO)" (SPP-1999). The authors of this work take full responsibility for its content.

References

David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. 1996. Active learning with statistical models. *CoRR*, cs.AI/9603104.

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

Liat Ein-Dor, Eyal Shnarch, Lena Dankin, Alon Halfon, Benjamin Sznajder, Ariel Gera, Carlos Alzate, Martin Gleize, Leshem Choshen, Yufang Hou, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. 2020. Corpus wide argument mining—a working solution. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7683–7691.

Yarin Gal. 2016. *Uncertainty in Deep Learning*. Ph.D. thesis, University of Cambridge.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA. PMLR.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep bayesian active learning with image data. *CoRR*, abs/1703.02910.

Ivan Habernal and Iryna Gurevych. 2016a. 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, Austin, Texas. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2016b. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany. Association for Computational Linguistics.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning.

Rong Hu. 2011. *Active Learning for Text Classification*. Doctoral Thesis, Technological University Dublin.

Anne Lauscher, Lily Ng, Courtney Napoles, and Joel Tetreault. 2020. Rhetoric, logic, and dialectic: Advancing theory-based argument quality assessment in natural language processing.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 543–552.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*.

Aditya Siddhant and Zachary C. Lipton. 2018. Deep bayesian active learning for natural language processing: Results of a large-scale empirical study. *CoRR*, abs/1808.05697.

Edwin Simpson and Iryna Gurevych. 2018. Finding convincing arguments using scalable Bayesian preference learning. *Transactions of the Association for Computational Linguistics*, 6:357–371.

Robin Smith. 2020. Aristotle's Logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Fall 2020 edition. Metaphysics Research Lab, Stanford University.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment – new datasets and methods.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

A Appendix

A.1 Uncertainty-based Acquisition Functions

In our work, we refer in particular to the following uncertainty-based acquisition functions (Gal et al., 2017):

- **variation ratios**: given a set of labels y_T from T stochastic forward passes, variation ratio for a given input point is calculated as:

$$varrat(x) = 1 - \frac{f_x}{T} \tag{1}$$

  with f_x denoting the number of times the most commonly occurring category (mode of the distribution) has been sampled. This serves as an indication of how concentrated the predictions are, with 0.5 being the highest dispersion, i.e. uncertainty, and 0 being the highest concentration (certainty) in the case of binary classification.

- **predictive entropy**: stems from information theory and is calculated by averaging the softmax values for each class :

$$predentr(x) = -\sum_c p(y = c|\mathbf{x}, D_{train}) \times \log_2(p(y = c|\mathbf{x}, D_{train})), \tag{2}$$

  where $p(y = c|\mathbf{x}, D_{train})$ stands for average probability of a data point adhering to a specific class given the outputs of the stochastic forward passes and the training data. c denotes the label class, i.e. we sum the values over all the classes to receive a measure of entropy for a given data point.

- **Bayesian Active Learning by Disagreement (BALD)** (Houlsby et al., 2011), also called **mutual information** (Gal, 2016), is a function of predictive entropy as described above and averaged predictive entropies that have been calculated separately for each output:

$$\begin{aligned} bald(x) = -[&\sum_c p(y = c|\mathbf{x}, D_{train}) \\ &\times \log(p(y = c|\mathbf{x}, D_{train}))] \\ + E_{p(\omega|D_{train})}[&\sum_c p(y = c|\mathbf{w}, \omega) \\ &\times \log(p(y = c|\mathbf{x}, \omega))]. \end{aligned} \tag{3}$$

A.2 Computing & Software Infrastructure

The experiments were conducted on a Ubuntu 18.04 system with an AMD Ryzen Processor with 16 CPU-Cores, 126 GB memory, and a single NVIDIA RTX 2080 GPU with 11 GB memory. We further used Python 3.7, PyTorch 1.4 and the Huggingface-Transformer library (2.11.0).

A.3 Topic Size and Difficulty

No.	Topic	Size	Acc.
0	Ban Plastic Water Bottles?	688	0.86
1	Christianity or Atheism	588	0.81
2	Evolution vs. Creation	782	0.78
3	Firefox vs. Internet Explorer	748	0.81
4	Gay marriage - right or wrong?	851	0.8
5	Should parents use spanking?	706	0.76
6	If your spouse committed murder, would you turn them in?	687	0.67
7	India has the potential to lead the world	822	0.81
8	Is it better to have a lousy father or to be fatherless?	616	0.64
9	Is porn wrong?	571	0.79
10	*Is the school uniform a good or bad idea?*	*878*	*0.78*
11	Pro choice vs. Pro life	845	0.61
12	Should physical edu. be mandatory?	568	0.74
13	*TV is better than books*	*747*	*0.79*
14	*Personal pursuit or common good?*	*733*	*0.84*
15	Farquhar as the founder of Singapore	820	0.7
	Total Size/Average Acc.	**11 650**	**0.76**

Table 4: Topic sizes in *UKPConvArg1Strict*. Topics are provided with their corresponding numbers and size within the data set, as well as our model's performance at test time. The topics selected for testing the acquisition functions have been highlighted in italics.

No.	Topic	Size	Acc.
0	Should flu vaccinations be mandatory?	731	0.75
1	Should gambling be banned?	503	0.8
2	Does online shopping bring more harm than good?	278	0.79
3	*Does social media bring more harm than good?*	*2587*	*0.78*
4	*Should we adopt cryptocurrency?*	*719*	*0.82*
5	Should we adopt vegetarianism?	1073	0.77
6	Should we sale violent video games to minors?	484	0.74
7	*Should we ban fossil fuels?*	*263*	*0.73*
8	Should we legalize doping in sport?	737	0.77
9	Should we limit autonomous cars?	1217	0.79
10	Should we support information privacy laws?	533	0.77
	Total Size/Average Acc.	**9 125**	**0.77**

Table 5: Topic sizes in *IBM-9.1kPairs*. Topics are provided with their corresponding numbers and size within the data set, as well as our model's performance at test time. The topics selected for testing the acquisition functions have been highlighted in italics.

A.4 Hyperparameters

For the evaluation we initialized all methods for **ten** runs with different seeds and reported the **mean accuracy score**. We used early stopping with a patience of three on a pre-selected validation set for regularization. As loss function we used weighted binary-cross-entropy for the (relative) Argument Strength task.

We train our models on top of the pre-trained BERT-Base uncased with a dropout probability of 0.1. Learning rate is 2^{-5} (same as in (Toledo et al., 2019)). The batch size per GPU is 64 and the model is validated after every half epoch.

7 Towards a Holistic View on Argument Quality Prediction

The chapter includes the **preprint** that is submitted and currently under peer-review:

> **Michael Fromm**, Max Berrendorf, Johanna Reiml, Isabelle Mayerhofer, Siddharth Bhargava, Evgeniy Faerman, and Thomas Seidl. „Towards a Holistic View on Argument Quality Prediction.“ In: Peer-Reviewing Phase

and the code is available at:

> `https://anonymous.4open.science/r/kdd-holistic-view-aq-C0D8`

Declaration of Authorship The research idea was proposed, developed and conceptualized by Michael Fromm. Michael Fromm did the design of the architecture and the pipeline. Johanna Reiml, Isabelle Mayerhofer, Siddharth Bhargava did the implementation and conducted the experiments. Michael Fromm, Evgeniy Faerman, and Max Berrendorf analyzed and discussed the results. Michael Fromm, Evgeniy Faerman, and Max Berrendorf wrote the manuscript. All authors revised the manuscript.

PREPRINT: Towards a Holistic View on Argument Quality Prediction

Michael Fromm[1], Max Berrendorf[1], Johanna Reiml[2], Isabelle Mayerhofer[2], Siddharth Bhargava[2], Evgeniy Faerman[1], Thomas Seidl[1]
[1] Database Systems and Data Mining, LMU Munich, Germany
[2] LMU Munich, Germany
{fromm, berrendorf, seidl}@dbs.ifi.lmu.de

ABSTRACT

Argumentation is one of society's foundational pillars, and, sparked by advances in NLP and the vast availability of text data, automated mining of arguments receives increasing attention. A decisive property of arguments is their strength or quality. While there are works on the automated estimation of argument strength, their scope is narrow: they focus on isolated datasets and neglect the interactions with related argument mining tasks, such as argument identification, evidence detection, or emotional appeal. In this work, we close this gap by approaching argument quality estimation from multiple different angles: Grounded on rich results from thorough empirical evaluations, we assess the generalization capabilities of argument quality estimation across diverse domains, the interplay with related argument mining tasks, and the impact of emotions on perceived argument strength. We find that generalization depends on a sufficient representation of different domains in the training part. In zero-shot transfer and multi-task experiments, we reveal that argument quality is among the more challenging tasks but can improve others. Finally, we show that emotions play a minor role in argument quality than is often assumed. We publish our code at https://anonymous.4open.science/r/kdd-holistic-view-aq-C0D8.

KEYWORDS

Argument Mining, Argument Quality, NLP

ACM Reference Format:
Michael Fromm[1], Max Berrendorf[1], Johanna Reiml[2], Isabelle Mayerhofer[2], Siddharth Bhargava[2], Evgeniy Faerman[1], Thomas Seidl[1]. 2022. PREPRINT: Towards a Holistic View on Argument Quality Prediction. In . , 10 pages.

1 INTRODUCTION

The argumentation process is one of the cornerstones of society, as it allows the exchange of opinions and reaching a consensus together. Fueled by advances in natural language processing, recent years have witnessed the advent of Argument Mining (AM), i.e., the field of automated discovery and organization of arguments. AM is helpful over various scenarios, reaching from legal reasoning [34, 48, 52, 54] to supporting the decision-making process of politicians [2, 10, 19, 24–26]. Thus, there is a flurry of works on identification of arguments from text [14, 40, 45] and retrieval of them [8, 9, 13, 39, 51]. Since arguments often have to be weighed against each other, a central property of arguments is their Argument Quality (AQ) or *convincingness*, i.e., their (perceived) strength. While the ancient Greeks [35] already discussed the constituents of strong arguments, automated estimation is a relatively uncharted field. Due to the high subjectivity of argument strength [16, 18, 40, 42, 43], obtaining high-quality annotations is challenging. In this light, a legitimate question is the reliability and robustness of the existing approaches for estimating AQ and their applicability in real-life scenarios. Existing AQ benchmark datasets are often restricted to a single domain [32, 49] or/and make different assumptions about factors impacting the AQ. Thus, enabling transfer between sources and datasets appears especially appealing, but existing works [16, 18, 42, 43] cease to provide detailed studies thereupon. Moreover, social science research suggests that the strength of an argument depends less on its logical coherence and depends much more appealing to the recipient's emotions [3, 4, 6, 20, 23] – a fact which is insufficiently considered so far.

Table 1: Two example arguments from the studied datasets with Argument Quality score and predicted emotion label and model confidence.

Topic: Polygamy Legalization	**Score**	**Emotionality**
"Polygamy makes for unhappy relationships and is patriarchal."	0.66	emotional (94.90%)
"Polygamy makes child-raising easier by spreading the needs of children across more people."	0.84	non-emotional (90.85%)

In this work, we thus investigate for the first time the automatic evaluation of the quality of arguments from a holistic perspective, bringing together various aspects: First, we evaluate whether AQ models can generalize across datasets and domains, which is a crucial feature for deployment in the diverse environments encountered in relevant real-world applications. Next, we investigate the hypothesis of whether models for related argument mining tasks inherently learn the concept of argument strength without being explicitly trained to do so by evaluating their zero-shot performance for estimating AQ. Finally, we investigate the effect of emotions in arguments: We present the first dataset for emotions in argumentative texts and demonstrate that emotions can be detected automatically therein, cf. Table 1. The obtained emotion detection models enable us to then provide evidence across all datasets examined that, contrary to the previous belief, emotional argumentation does not significantly influence perceived argument strength.

In summary, our contributions are as follows:

- We are the first to study the generalization capabilities of AQ prediction models across different datasets and AQ notions.
- Since we determine the size of the dataset as one of the decisive performance factors, we further investigate a zero-shot setting of transferring from related Argument Mining tasks.
- Finally, we elucidate the relation between emotions and AQ. To this end, we provide a novel dataset for emotion in argumentative texts and show that these can be predicted on par with human performance. Using this capable emotion detection model, we then show that, in contrast to popular belief, the AQ of emotional arguments does *not significantly differ* from non-emotional ones, at least across four different publicly available AQ datasets.

2 RELATED WORK

2.1 Argument Quality

Argument Quality (AQ), sometimes also called Argument Strength, is a sub-task in Argument Mining (AM) that belongs to the central research topics among argumentation scholars [44, 46, 53]. Due to its high subjectivity, there is no single definition of AQ. Therefore, there are various suggestions on different factors that can affect an argument's quality, e.g., the *convincingness* of an argument [17]. To the best of our knowledge, we are the first who evaluate how these factors correlate with each other across different corpora. Furthermore, there are various possibilities to express the strength of an argument. Some works adopt an absolute continuous score, while others advocate that strength estimation works better in (pairwise) relation to other arguments.

One of the first relatively large corpora was introduced by Swanson et al. [42]. The SwanRank corpus contains over 5k arguments, where each argument is labeled with a continuous score that describes the interpretability of an argument in the context of a topic. They propose a method using linear regression, ordinary Kriging, and SVMs as regression algorithms to estimate the strength automatically from an input text encoding by handcrafted features. Other corpora followed and used the relative- and/or absolute *convincingness* [18, 33] as the annotation criterion. The authors proposed models based on SVMs or BiLSTM combined with GloVe embeddings [31]. Gleize et al. [15] provide a dataset, *IBM-EviConv*, focused on ranking the *evidence* convincingness. They used a Siamese network based on a BiLSTM with attention and trainable Word2Vec embeddings. Gretz et al. [16] and Toledo et al. [43] created their corpora by asking annotators if they would recommend a friend to use the argument in a speech supporting/contesting the topic, regardless of their own opinion. Both use a fine-tuned BERT [7] model for the absolute AQ regression task.

The shared evaluation practice in the previous works is to evaluate methods on each dataset independently. Gretz et al. [16] use the newly introduced dataset for model pre-training but then fine-tune the model on the training part of the dataset used for the evaluation. This work proposes to advance evaluation and advocate for an accurate *cross-dataset* evaluation without additional fine-tuning on the evaluation dataset to estimate the model's applicability in challenging *real-life* scenarios.

2.2 Role of Emotions in Argumentation

The previous works only empirically investigate the role of emotions in argumentation on a small scale. Wachsmuth et al. [50] created a corpus of 320 arguments, annotated for 15 fine-grained argument dimensions originating from argument theory. They categorize the quality dimensions into three main quality aspects: *Logical*, *rhetorical*, or *dialectical* quality. One dimension in rhetorical quality is the *emotional appeal*, defined as: *"Argumentation makes a successful emotional appeal if it creates emotions in a way that makes the target audience more open to the author's arguments"*. The authors did not find any significant correlations to other quality dimensions.

Benlamine et al. [3, 4] showed in an experimental setting with 20 participants that the mode of *pathos* represented by emotion is essential in the persuasion process in argumentation. Their experiment indicates that *"[the] Pathos strategy is the most effective to use in argumentation and to convince the participants"*.

In both works, the sample size is relatively small (20 participants or 320 arguments). To better substantiate the considerations, we investigate the influence of emotional appeals regarding AQ annotations on more than 40k arguments across four large corpora.

3 GENERALIZATION ACROSS ARGUMENT QUALITY CORPORA

High-level applications such as Argument Retrieval [8, 9, 13, 39, 51] and autonomous debating systems [38] require reliable Argument Quality (AQ) models to select strong arguments among the relevant ones. The research community has identified this gap and proposed and evaluated different automated models for AQ estimation [16, 18, 42, 43]. However, AQ is often captured differently due to its high subjectivity, e.g., absolutely as a continuous score or relative to other arguments by pairwise comparison. Consequently, many publications also introduced their corpus with individual annotation schemes capturing different notions of AQ. While they compared multiple models against each other *within* a single corpus, there is a lack of *cross-corpora* empirical evaluations. Thus, the robustness of predictions across datasets remains largely unexplored, which poses a severe challenge for reliable real-world applications integrating diverse data sources. To evaluate the generalization capability of AQ estimation models, we designed a set of experiments across all four major AQ datasets to answer the following research questions:

- How well do AQ models perform across datasets if annotations schema and domain of the arguments do not change?
- How does the corpora size affect generalization?
- How well do models generalize across different text domains?
- How does the AQ quality notion affect generalization?
- Does the AQ model become more robust if it is trained with a combined dataset containing data from different domains and labeling assumptions also vary?

3.1 Evaluation Setting

We briefly describe the four AQ datasets used in our empirical study, which all capture AQ on a sentence level. They are also summarized in Table 2. Swanson et al. [42] constructed the dataset SwanRank

Table 2: Overview of the different Argument Quality (AQ) datasets with their number of arguments, the number of distinct topics, the different source domains, and the AQ notion used for annotation.

Name	Sentences	Topics	Domain	Quality notion
UKPConvArgRank [18]	1,052	32	Debate Portal	Convincingness
SwanRank [42]	5,375	4	Debate Portal	Interpretability
IBM-ArgQ [43]	5,300	11	Crowd Collection	Recommendableness
IBM-Rank [16]	30,497	71	Crowd Collection	Recommendableness

with over 5k arguments whose quality is labeled in the range of $[0, 1]$, where 1 indicates that an argument can be *easily interpreted*. Habernal et al. [18] annotated a large corpus of 16k argument pairs and investigated which argument from the pair is more *convincing*. Based on the argument pair annotations, they created an argument graph and used PageRank to calculate absolute scores for the individual arguments. The result is called `UKPConvArgRank` and contains 1k arguments. Gretz et al. [16] and Toledo et al. [43] created their corpora of 30k and 6.3k arguments by asking annotators if they would *recommend a friend to use the argument* in a speech supporting or contesting the topic regardless of their personal opinion. Gretz et al. [16] used crowd contributors that presumably better represent the general population, compared to debate club members that annotated in Toledo et al. [43]. Furthermore, Gretz et al. [16] also considered the annotators' credibility without removing them entirely from the labeled data, as done in Toledo et al. [43].

As some of the corpora did not provide official train-validation-test splits and differed in the number of topics and the formulated task (in-topic vs. cross-topic), we decided to do our own split based on the topics of the arguments. We perform 10-fold cross-topic cross-validation, where each fold is a 60%/20%/20% train-validation-test split, and we additionally ensure that no topic occurs in more than one split. By the latter requirement, we ensure an inductive setting where the AQ estimation can not rely on similar arguments in the training corpus and therefore provides a more challenging but more realistic task.

3.2 Model and Training

Since transfer learning achieves state-of-the-art Argument Mining (AM) results on different corpora and tasks [14, 36, 45], we also apply it to our AQ estimation task. We use a bert-base model, pre-trained on masked-language-modeling, and fine-tune it to predict absolute AQ scores on the respective datasets, cf. Section 3.1. As an input, we used the arguments from the respective datasets and concatenated the topic information, separated by the BERT specific $[SEP]$ limiter, similar to other work in argument mining [14, 16, 36]. We concatenate the last four layers of the fine-tuned BERT model output to obtain an embedding vector of the size $4 \cdot 768 = 3,072$. For the regression task, we stack a Multi-Layer Perceptron (MLP) with two hidden layers, one with 100 neurons and a ReLU activation, followed by the second hidden layer and a sigmoid activation function. We train the architecture end-to-end, with SGD with a weight decay of 0.35 and a learning rate of $9.1 \cdot 10^{-6}$. The MLP uses dropout with a rate of 10%.

3.3 Results

Table 3 summarizes our results. We report the Pearson correlation score between the predicted- and ground-truth absolute AQ evaluated on a hold-out test set.

3.3.1 Evaluation on Similar Datasets and Importance of Training Set Size. First, we evaluate the performance of the model on similar datasets and the dependency on the size of the training dataset. We can observe that models perform very well on other datasets from the same domain labeled with a similar quality notion, i.e., IBM-ArgQ and IBM-Rank datasets. Furthermore, we can notice that the size of the dataset is crucial for performance: a model trained on the largest IBM-Rank dataset achieves the best score also on IBM-ArgQ. This insight gives us a solid foundation for the next steps.

3.3.2 Generalization Across Domains and Quality Notions. Next, we investigate whether a transfer across domains is possible. To this end, we train on one dataset and evaluate on a different one. Recall that the four datasets cover two different domains: the sentences from UKPConvArgRank and SwanRank have been extracted from debate portals, while IBM-Rank and IBM-ArgQ have been collected from the crowd.

Compared to in-domain generalization, we observe a considerably worse generalization between domains: For example, trained on the crowd dataset IBM-ArgQ, we can achieve a correlation of 38.9% on the crowd dataset IBM-Rank, while training on the debate datasets SwanRank and UKPConvArgRank results in negligibly low correlations of 8% and 3%, respectively. Conversely, when evaluated on the debate portal dataset SwanRank, we obtain a correlation of 42.5% when using a model trained on the other debate portal dataset UKPConvArgRank, while the crowd collected datasets IBM-ArgQ and IBM-Rank only achieves 27.8% and 37.0%, respectively. The smaller difference compared to the first comparison can be explained by the larger size of the training datasets.

Surprisingly, we observe a completely different picture for generalization across quality notions. We see only a moderate drop in performance for a fixed domain but a different quality notion. For instance, the model trained on SwanRank performs relatively well on the UKPConvArgRank dataset. Vice-versa, we observe a more considerable performance drop, which can be explained by the smaller size of the UKPConvArgRank dataset.

3.3.3 Multi-Domain and Multi-Quality Notion Training. To investigate whether a single model can grasp various dimensions of quality and work on arguments from various domains, we designed another set of "leave-one-out" experiments. We train on the training

Table 3: The models are evaluated by the Pearson correlation between ground truth and predicted Argument Quality on the respective test sets. The first four rows correspond to models trained on a single dataset, whereas for the last four rows, *all but one* dataset, have been used for training, i.e., following a leave-one-out scheme. Bold numbers indicate the best results for each column within the two groups.

			Evaluation			
		Size	UKPConvArgRank	SwanRank	IBM-ArgQ	IBM-Rank
Training	UKPConvArgRank	1,052	19.0%	42.5%	15.2%	3.0%
	SwanRank	5,375	18.9%	**47.5%**	17.1%	8,0%
	IBM-ArgQ	5,300	23.3%	27.8%	34.2%	38.9%
	IBM-Rank	30,497	**26.2%**	37.0%	**38.3%**	**48.1%**
	all except UKPConvArgRank	41,172	23.3%	45.8%	31.6%	46.6%
	all except SwanRank	36,849	**25.0%**	**49.1%**	35.0%	46.6%
	all except IBM-ArgQ	36.924	23.0%	43.6%	**38.4%**	**47.5%**
	all except IBM-Rank	12.224	20.4%	42.0%	35.0%	46.5%

sentences of all but one AQ corpus and evaluate the performance on all test sets. The entries on the diagonal thus show how well the models perform when evaluated on an unseen corpus.

For evaluation on the unseen IBM-Rank dataset after training on the remaining ones, we can obtain a correlation of 46.5%, which nearly reaches the correlation of 48.1% we obtained when training and evaluating on IBM-Rank. For SwanRank, IBM-ArgQ and UKPConvArgRank, we can even surpass the correlation on the respective test set by training on all other training sets instead of the one from the respective corpus.

3.3.4 Cross-Corpora Generalization Conclusion. To summarize, we conclude that, in general, the available datasets and models for AQ are reliable, and the models can grasp the concepts automatically. Our most important insight is that AQ notions do not contradict each other, and a single model can estimate the AQ of text from different domains. Therefore, the practical recommendation for real-life application is to combine all available datasets across different domains and AQ notions.

4 ZERO-SHOT-LEARNING IN ARGUMENT MINING

In this section, we investigate whether explicit Argument Quality (AQ) corpora are a necessity, or whether the task of AQ can also be solved by transferring from other related argument mining tasks such as Argument Identification (AId) or Evidence Detection (ED), In contrast to the relatively new task of automatic AQ estimation, other Argument Mining (AM) tasks already offer a broad range of large datasets that cover different domains and annotation schemes. Moreover, the agreement between the annotators is higher on the other tasks, as AQ is highly subjective [16, 18, 40, 42, 43]. Therefore, a successful transfer from related tasks to the target task of AQ would represent a significant advance in the field. To this end, we investigate the zero-shot capability of AM models across different corpora *and* different AM tasks. To the best of our knowledge, we are the first to compare AM task similarity by providing a first study on how individual tasks can benefit from each other.

In particular, we aim to answer the following guiding research questions:

- Can we achieve satisfactory performance by zero-shot transfer from related AM tasks, i.e., without fine-tuning the respective task?
- Is there a difference in transferring from different tasks, i.e., is one task more suited than the other?

While not a primary focus of this work, for completeness, we also provide experimental results for the reverse direction of transferring *from* AQ estimation *to* the other tasks.

4.1 Datasets

Table 4 provides an overview of the different AM corpora we used in our experiments, covering three different AM tasks. UKP-Sentential [40] contains over 25k arguments distributed across eight controversial topics. It is annotated for AId, where each argument is labeled as either *argumentative* or *non-argumentative* in the context of a topic. The IBM-Evidence [12] corpus includes nearly 30k sentences from Wikipedia articles. All sentences are annotated with a score in the range of $[0, 1]$, denoting the confidence that the sentence is evidence (either expert or study evidence) to the article's topic. IBM-Rank [16] is the largest of the four AQ datasets, which has also been used in the previous Section 3. The corpus' annotation is in the range of $[0, 1]$, where 1 indicates a strong argument and a score of 0 indicates a weak argument. We split all three datasets into train, validation, and test sets (70%/10%/20%). Similar to Section 3.1, we designed the splits such that no topic in the training set also occurs in the test set, which is often called the "cross-topic" scenario in AM and corresponds to a more interesting, but also more challenging task, which requires a sufficient degree of generalization to unseen topics.

4.2 Evaluation Setting

We use a standard BERT large model [7] pre-trained on the masked-language-modeling task to evaluate the zero-shot generalization capability. As an input for the fine-tuning, we use the sentences from the respective datasets and concatenate the topic information, separated by the BERT specific `[SEP]` limiter, similar to Section 3.2. We develop three different zero-shot evaluation strategies for the different transfer settings:

Table 4: Overview of the different Argument Mining (AM) datasets, we used for the zero-shot experiments, with their size in terms of the number of sentences, the number of covered topics, the source domain and the AM task.

Name	Sentences	Topics	Domain	Task
IBM-Rank [16]	30,497	71	Crowd Collection	Argument Quality (AQ)
UKP-Sentential [40]	25,492	8	Web Documents	Argument Identification (AId)
IBM-Evidence [12]	29,429	221	Wikipedia	Evidence Detection (ED)

Table 5: Zero-Shot performance of the Argument Mining models. The evaluation measure is Macro F_1 for Argument Identification (AId), and the Spearman correlation for Evidence Detection (ED) and Argument Quality (AQ).

Train	Evaluation		
	AId	ED	AQ
AId	73.53% ± 3.21%	53.80% ± 1.23%	25.72% ± 1.31%
ED	75.17% ± 0.70%	77.90% ± 0.23%	28.66% ± 0.92%
AQ	71.27% ± 0.74%	43.51% ± 3.10%	47.45% ± 1.16%
Metric:	Macro F_1	ρ	ρ

- **AId → Regression Tasks**: We use the BERT encoder output as input to a linear layer with dropout that predicts the classes. Cross-entropy serves as training loss. The probabilities between 0 and 1 indicate if a sentence is argumentative or not. The predicted probability of the positive class, i.e., whether it is argumentative, is then directly used as a score for ED and AQ on the respective corpora. We use Spearman rank-correlation instead of Pearson correlation as an evaluation measure to account for the difference in scale.
- **Regression Tasks → AId**: ED and AQ use the BERT representations in a single hidden layer that scores the sentences according to their absolute quality or the probability of containing evidence. Since we train on regression tasks, we use the Mean Squared Error loss during training. We then apply the trained models to AId. We select an optimal decision threshold α among all possible thresholds on UKP-Sentential's validation set according to Macro F_1. This model is then evaluated on the UKP-Sentential test set.
- **Regression Task ↔ Regression Task**: For the evaluation between two regressions models, we calculate the Spearman correlation coefficient directly on their respective outputs.

4.3 Results

Table 5 shows the results from our experiments. We train three models with different random seeds for each training task and report the mean and standard deviation of evaluation on the different tasks.

We generally observe, unsurprisingly, that training on the same task as evaluating yields the best results with Spearman correlations of ≈ 77.90% for ED → ED and ≈ 47.45% for AQ → AQ.

A notable exception is AId, where a model trained on ED achieves ≈ 75.17% Macro F_1 and thus can slightly surpass the performance of a model directly trained on AId of ≈ 73.53%, although within the range of one standard deviation. Exceeding the in-task performance is a strong result, as the model has never explicitly been trained for the task. We generally observe almost perfect zero-shot transfer towards AId, as also the model trained on AQ achieves a performance of ≈ 71.27%, which is only 2% points behind the ≈ 73.53% from AId to AId. Thus, models capable of predicting whether a sentence provides evidence (ED) or capable of predicting the AQ of an argument, inherently learn concepts that enable the detection of whether a sentence is argumentative or not (AId). To further give context to the zero-shot performance, the BiCLSTM approach trained on the AId task from [40] obtained a Macro F_1 of 64.14%, i.e., worse results than the zero-shot transfer despite explicitly being trained on the task, which underlines the remarkable zero-shot performance, and may indicate that AId is a simpler task than the other two, ED and AQ.

For ED, we achieve the best performance of ≈ 77.90% Spearman correlation by directly training on this task. The model trained on AId obtains the closest zero-shot transfer result with a rank correlation of ≈ 53.80%, which still represents a considerable correlation, despite being ≈ 24% points behind. The model trained for AQ shows the worst transfer from the studied tasks with a correlation of ≈ 43.51%. Overall, we note that the challenging zero-shot transfer is still possible with an acceptable loss in performance. Models trained on detecting whether a sentence is argumentative or not (AId) transfer better than those trained for predicting the argumentative strength of a sentence AQ to the target task of predicting the confidence whether a sentence provides evidence (ED).

For AQ, the main focus of our paper, we achieve the best performance of ≈ 47.45% Spearman correlation by directly training on this task. When transferring from related AM tasks in a zero-shot setting, we have to tolerate decreases in performance to ≈ 28.66% for transfer from ED, and ≈ 25.72% for transfer from AId, respectively. Thus, models capable of detecting whether a sentence is argumentative (AId) are slightly less well applicable to predicting the sentence's argumentative strength than the models for predicting a level of supporting evidence (ED). One factor here may be that ED is also a regression task as opposed to the classification task of AId.

To summarize, the results suggest that the tasks of AId, i.e., classifying whether a sentence is argumentative, and ED, i.e., predicting a numeric level of supporting evidence, are closer to each other than to the more difficult task of assessing the argumentative strength, as witnessed by worse zero-shot transfer results from and to AQ. Nevertheless, in principle, a transfer in the highly challenging zero-shot setting is possible; for closer related tasks, it can even lead to similar scores as training directly on the target task.

Table 6: Performance of multi-task models trained on different Argument Mining task combinations, including Argument Identification (AId) and Evidence Detection (ED). The performance is measured by Macro F_1 for AId, and the Spearman correlation for ED and AQ.

Train	Evaluation		
	AId	ED	AQ
AQ	-	-	$47.45\% \pm 1.16\%$
AQ/AId	$80.07\% \pm 1.16\%$	-	$47.46\% \pm 0.58\%$
AQ/ED	-	$78.07\% \pm 0.45\%$	$46.84\% \pm 0.25\%$
AQ/AId/ED	$78.91\% \pm 3.17\%$	$78.40\% \pm 0.03\%$	$48.39\% \pm 1.12\%$
Metric:	Macro F_1	ρ	ρ

Table 7: Overview of the different emotion detection datasets from heterogeneous text domains used for the behavioral fine-tuning of EmoBERT.

Name	Sentences	Domain
Alm [1]	15,036	Childrens' stories
ISEAR [37]	7,666	Reactions and emotion antecedents
SemEval-2007 [41]	1,250	News headlines
SemEval-2018 [27]	9,625	Tweets
SemEval-2019 [5]	14,335	Dialogues
Neviarouskaya 2010 [29]	1,000	Stories
Neviarouskaya 2011 [30]	700	Diary-like-blogs

4.4 Multi-Task Learning for Argument Quality

As shown in the last section, the AM tasks are sufficiently close to each other to enable successful zero-shot transfer. An interesting question that arises from this observation is whether the performance in AQ estimation further improves by multi-task learning. To this end, we developed a multi-task model that involves a shared BERT encoder and separated linear layers for the respective tasks. We trained the architecture with weighted loss functions, ensuring that each task is weighted equally. Our results are shown in Table 6. Focusing on the right-most column first, we can see that the performance in terms of Spearman correlation only marginally improves by multi-task learning. A possible explanation is here that we already observed that the other two tasks are seemingly less challenging and more closely related to each other than to AQ. As additional supporting evidence, ED slightly and AId considerably benefits from multi-task learning with AQ.

5 EMOTION DETECTION

Most work in Argument Mining (AM) focuses on the *logos* mode of persuasion, i.e., whether arguments are logically plausible. Nevertheless, recent studies support that the mode of *pathos* represented by emotions is essential in the persuasion process [3, 4, 6, 20, 23]. Those studies have in common that they relied on relatively small sample sizes. In the following, we thus evaluate the hypothesis that the AQ scores in the publicly available Argument Quality (AQ) datasets Swanson [42], UKP [18], Gretz [16], and Toledo [43], are influenced by appealing to the emotions of the annotators.

AQ datasets do not provide emotion labels, and therefore, we first need a reliable and scalable method to estimate the level of emotionality in the arguments. As to the best of our knowledge, there is no previous work on automatic emotion detection in arguments, we investigate various approaches from the very simple baselines to complex multi-step transfer learning models. For the evaluation and comparison of different methods, we create a novel argument dataset EmoArg-523, where for each argument, we manually annotate emotionality.

After the reliable emotion detection model is available, we apply it on the unlabeled arguments from the four AQ corpora to obtain proxy emotion labels. We then use these proxy labels to investigate the relation between emotions and argument AQ at a large scale.

In particular, we address the following research questions:

- Can we automatically detect emotions in argumentative texts from different domains?
- Can we substantiate the hypothesis that arguments arousing emotions are perceived stronger?

5.1 Datasets

5.1.1 Novel Emotional Argumentation Dataset. For the evaluation of our EmoBERT model, we sample 150 arguments from each of the four AQ corpora (IBM-Rank , IBM-ArgQ, SwanRank, UKPConvArgRank), i.e., 600 in total. These arguments are manually labeled by six independent annotators. The arguments were labeled based on the annotation guidelines as *emotional*, when the arguments contained pathos rhetoric, or *non-emotional* when the persuasion process in the argument was driven by evidence or logical rhetoric. The annotator agreement calculated via Krippendorff's Alpha [22] is 31.28%. Note that because of the subjectivity of the task, such an agreement is acceptable; for comparison, e.g., Wachsmuth et al. [50] achieved an Alpha of 26% for the quality dimension "Emotional appeal". After the agreement calculation, we removed the 77 sentences (12.8%) without a majority between the six annotators. The resulting dataset comprises 225 emotional (43.02%) and 298 (56.98%) non-emotional arguments and is referred to as EmoArg-523. We split the dataset into train-, validation-, and test-set (60%/10%/30%).

5.1.2 General Emotion Detection Dataset. Although it is not clear a priori that emotion detection transfers well from other domains to the domain of argumentation, we hypothesize that the model can benefit from existing datasets. Motivated by our results for AQ detection, where the model trained on a joined dataset demonstrated very robust performance, we combine seven emotion datasets [1, 5, 27, 29, 30, 37, 41], c.f., Table 7. The emotion datasets came with different classes of emotions. Thus, we unified these different label formats by assigning the existing labels for emotions, such as *happy*, *sad* or *fear*, or *neutral*, to a binary label of either *emotional* or *non-emotional*. The seven datasets were then split individually into train-, validation-, and test-set and combined to a large heterogeneous emotion corpus.

5.2 Models & Baselines

Since transfer learning achieves state-of-the-art results for AM on different corpora and tasks [14, 36, 45], we also apply it for the task of emotion detection. We employ a transformer [47] based BERT model [7] with fine-tuning on different datasets. As a regularization technique to avoid over-fitting, early stopping is used on the validation cross-entropy loss, with a patience value of three epochs. We include the following model variants and baselines in our evaluation:

Majority Baseline The majority baseline labels the arguments with the most frequent class based on our EmoArg-523 corpus, which is *non-emotional* (57.55%).

Pronouns Baseline The pronouns baseline labels the arguments as emotional, which contain at least one of the personal pronouns "I", "you" or "me".

NRC Baseline The NRC baseline labels the arguments which contain at least *one* unigram contained in the NRC Emotion Lexicon [28].

EmoBERT To assess the generalization, we evaluate the zero-shot performance of a BERT model fine-tuned on the heterogeneous emotion corpora, cf. see Section 5.1.2. The combined emotion corpora incorporate multiple domains found on the internet, and therefore, the resulting model is supposed to be universally applicable.

ArgBERT Bert-Base fine-tune it on 339 sentences of our annotated argument emotion dataset.

ArgBERT-EmoInit It is the same as ArgBERT-EmoInit, but we also fine-tune it on 339 sentences of our annotated argument emotion dataset. We hypothesize that with the two-step transfer learning approach, the model first learns a general concept of emotions and then can focus on the target argument domain.

Human Performance An interesting experiment for assessing the applicability of the proposed solution is the comparison with the human performance on the task. To compute the human performance, we evaluate each annotator against the majority label of the remaining annotators using the Macro F_1 score.

5.3 Results

5.3.1 Emotion Detection. We present the results in terms of Macro F_1 on the novel dataset EmoArg-523 for emotion detection in argumentative texts in Table 8. Despite its simplicity, the strongest baseline with a Macro F_1 score of 59.7% is the Pronouns Baseline (Pronouns Baseline), EmoBERT achieves a Macro F_1 of $\approx$ 67.1%, which highlights that *domain adoption* from the source - the heterogeneous emotion detection datasets - to the target domain of arguments *is possible*. The best emotion detection model, ArgBERT-EmoInit, which used behavioral fine-tuning on the emotion dataset, followed by a second fine-tuning on the dataset of emotion-annotated arguments, achieves a Macro F_1 score of $\approx$ 74.6%, only a few points below the human performance estimate of $\approx$ 80.9%. For most models, we also observe a slight decrease in performance between the test part and the full EmoArg-523 (EmoArg-523); we attribute this to a slight distribution shift where the test part seems to contain slightly more arguments with difficult to detect emotions.

Table 8: Overview of the emotion detection results for different model variants on the annotated Argument Quality dataset, EmoArg-523. We show results in terms of Macro F_1 for different BERT model variants, as well as the three baselines in addition to a human performance estimate. For those models which do not make use of the labels on EmoArg-523, we also report the performance across all labeled arguments. In bold font, we highlight the best performance inside one group.

Method	Split	
	train+val+test	test
Majority Baseline	36.3%	36.4%
Pronouns Baseline	**63.0%**	**59.7%**
NRC Baseline	52.3%	50.3%
EmoBERT	**67.1**% ± **3.0**%	65.9% ± 5.3%
ArgBERT	-	73.2% ± 1.8%
ArgBERT-EmoInit	-	**74.6**% ± **1.7**%
Human Performance	82.1% ± 4.0%	80.9% ± 4.0%

5.3.2 The Effect of Emotions on Argument Quality. We start by analyzing the relation of emotionally appealing texts and AQ on the relatively small test part of the novel annotated dataset, EmoArg-523. Fig. 1 shows the distribution of AQ for emotional vs. non-emotional arguments based on the three different emotion detection models and the ground truth annotation grouped by dataset. Except for IBM-ArgQ, we observe the mean AQ of emotional arguments to be higher than those of non-emotional arguments. A possible explanation is that, in contrast to the other datasets that used crowd workers, the annotation on IBM-ArgQ was created by debate club members, who may have been trained to judge explicitly not considering an emotional appeal. However, partially due to the small sample size, the differences are insignificant ($p > 0.01$) according to Welsh's unequal variance t-test with Fischer adjustments.

Next, we utilize the trained emotion detection models to extend the analyses from the 157 test sentences in EmoArg-523 to the remaining 41,905 from the combined AQ corpora. While we are now restricted to predicted emotionality only instead of human annotations, we reviewed its quality in the previous section and found it sufficient. Fig. 2 shows the distribution of AQ grouped by predicted appeal to emotion for all three models and four datasets.

For SwanRank, emotional arguments receive slightly larger quality scores. While this is consistent across all models, it is clearly visible for EmoBERT and the most reliable emotion prediction model, ArgBERT-EmoInit. We attribute this to the covered topics of Gun Control, Gay Marriage, Death Penalty, and Evolution, which are areas with emotional discussions. On IBM-ArgQ, the differences are smaller but consistent across all models, with a slight tendency towards non-emotional arguments being perceived stronger. A possible explanation can be the annotation process, where debate club members served as annotators, which may be taught towards looking at logical arguments without letting emotions affect their view. The other two datasets do not show a consistent nor noticeable difference in the distributions. Overall, we cannot observe a clear

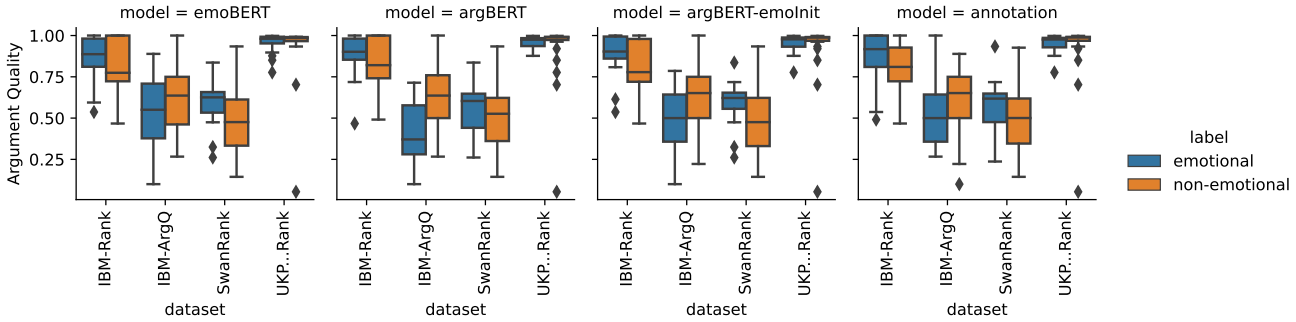

Figure 1: Relation between predicted / annotated emotionality and Argument Quality grouped by dataset for the three different models and the ground truth (on the right-most panel; denoted by *annotation*).

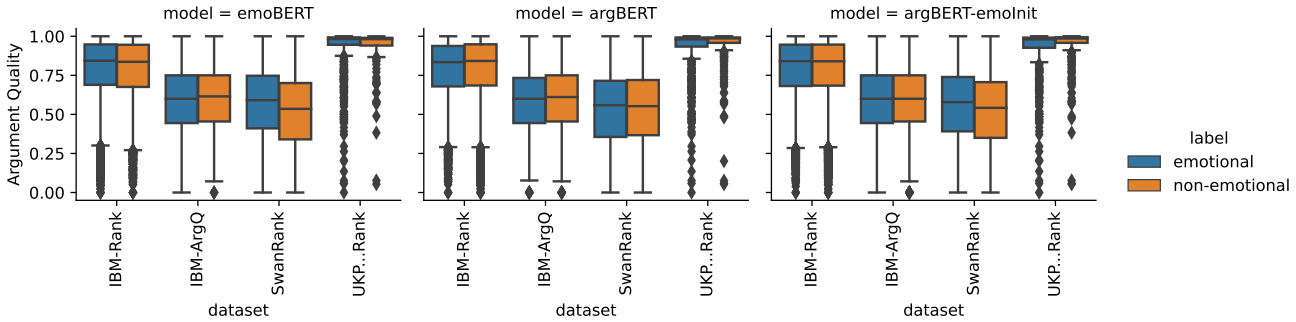

Figure 2: Comparison of the Argument Quality (AQ) of the remaining unlabeled $> 40,000$ arguments grouped by predicted emotionality across the four datasets.

relation between emotional argumentation and perceived strength on a large scale, challenging existing views. With our new dataset for emotional argumentation and the proxy models capable of predicting emotional argumentation, we hope to enable social sciences to study the deeper reasons behind this in the future.

6 CONCLUSION

We see this work as a fundamental step towards a holistic view of Argument Quality (AQ): We showed that for good generalization across individual AQ corpora, a match between the source and target domain of the arguments is essential. In contrast, diversity in AQ notions does not hinder but rather enriches the generalization capability. The target domain has a minor impact with sufficient broad coverage of different domains and adequate size. This insight is directly actionable for practical applications: The benefits of different AQ notions permit direct integration of different data sources, which is a prerequisite for dealing with the inputs from diverse domains encountered, e.g., by general-purpose argument retrieval engines.

Moreover, we could elucidate AQ's relation to other Argument Mining (AM) tasks, such as Evidence Detection (ED) and Argument Identification (AId). Our zero-shot transfer experiments demonstrated that the concepts learned for one of the tasks are sufficient to solve the other to some degree without explicitly being trained for it. By comparing the achieved results, we conclude that AId and ED are more closely related to each other than to AQ, and per se also easier to transfer to it. The multi-task experiment further emphasized this, where AQ could gain less from the other tasks than vice-versa. Thus, an important open question is how to enable better successful transfer towards AQ, and also extending beyond the three tasks we studied in this work.

Finally, we provide the community with a new corpus that consists of AQ *and* emotion annotations. In contrast to some results from social science research, our extensive empirical evaluation across a large number of argumentative sentences found overall only a limited influence of emotional appeal on the AQ scores. A deeper analysis of these surprising results' (social) determinants is an important future work direction. Besides the well-studied *logos* dimension of logical plausibility and the *pathos* dimension investigated in this work, the third remaining dimension from classical argumentation theory is *ethos*, which did not receive sufficient attention by the AM community so far. Existing smaller datasets [11, 21] invite to visit these uncharted territories, e.g., by studying argument strength in context to the provenance of an argument.

REFERENCES

[1] Cecilia Ovesdotter Alm. 2009. Affect in Text and Speech.

[2] Rawia Awadallah, Maya Ramanath, and Gerhard Weikum. 2012. Harmony and Dissonance: Organizing the People's Voices on Political Controversies. In *Proc. of the Fifth ACM Int. Conf. on Web Search and Data Mining* (Seattle, Washington, USA) *(WSDM '12)*. 523–532.

[3] Mohamed S. Benlamine, Serena Villata, Ramla Ghali, Claude Frasson, Fabien Gandon, and Elena Cabrio. 2017. Persuasive Argumentation and Emotions: An Empirical Evaluation with Users. In *Human-Computer Interaction. User Interface Design, Development and Multimodality*. 659–671.

[4] Sahbi Benlamine, Maher Chaouachi, Serena Villata, Elena Cabrio, Claude Frasson, and Fabien Gandon. 2015. Emotions in argumentation: an empirical evaluation. In *Proc. of the 24th Int. Conf. on Artificial Intelligence*. 156–163.

[5] Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. SemEval-2019 Task 3: EmoContext Contextual Emotion Detection in Text. In *Proc. of the 13th Int. Workshop on Semantic Evaluation*. 39–48.

[6] Francesca D'Errico, Marinella Paciello, and Matteo Amadei. 2018. Behind Our Words: Psychological Paths Underlying the Un/Supportive Stance Toward Immigrants in Social Media. In *DSAA*. 649–656.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*. 4171–4186.

[8] Lorik Dumani, Patrick J. Neumann, and Ralf Schenkel. 2020. A Framework for Argument Retrieval. In *Advances in IR*. 431–445.

[9] Lorik Dumani and Ralf Schenkel. 2019. A systematic comparison of methods for finding good premises for claims. In *Proc. of the 42nd Int. SIGIR*. 957–960.

[10] Rory Duthie, Katarzyna Budzynska, and Chris Reed. 2016. *Mining Ethos in Political Debate*. Frontiers in Artificial Intelligence and Applications, Vol. 287. 299–310.

[11] Rory Duthie, Katarzyna Budzynska, and Chris Reed. 2016. Mining ethos in political debate. In *Computational Models of Argument: Proc. from the Sixth Int. Conference on Computational Models of Argument (COMMA)*. 299–310.

[12] Liat Ein-Dor, Eyal Shnarch, Lena Dankin, Alon Halfon, Benjamin Sznajder, Ariel Gera, Carlos Alzate, Martin Gleize, Leshem Choshen, Yufang Hou, et al. 2020. Corpus wide argument mining—a working solution. In *Proc. of the AAAI Conf. on Artificial Intelligence*, Vol. 34. 7683–7691.

[13] Michael Fromm, Max Berrendorf, Sandra Obermeier, Thomas Seidl, and Evgeniy Faerman. 2021. Diversity Aware Relevance Learning for Argument Search. In *Advances in IR - 43rd European Conf. on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proc., Part II*, Vol. 12657. 264–271.

[14] Michael Fromm, Evgeniy Faerman, and Thomas Seidl. 2019. TACAM: Topic And Context Aware Argument Mining. In *2019 IEEE/WIC/ACM Int. Conf. on Web Intelligence, WI 2019, Thessaloniki, Greece, October 14-17, 2019*. 99–106.

[15] Martin Gleize, Eyal Shnarch, Leshem Choshen, Lena Dankin, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. 2019. Are You Convinced? Choosing the More Convincing Evidence with a Siamese Network. In *ACL*. 967–976.

[16] Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A Large-Scale Dataset for Argument Quality Ranking: Construction and Analysis. *AAAI Conf* 34, 05 (Apr. 2020), 7805–7813.

[17] Ivan Habernal and Iryna Gurevych. 2016. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *EMNLP*. 1214–1223.

[18] Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM. In *ACL*. 1589–1599.

[19] Shohreh Haddadan, Elena Cabrio, and Serena Villata. 2019. Yes, we can! Mining Arguments in 50 Years of US Presidential Campaign Debates. In *Proc. of the ACL*. 4684–4690.

[20] Denis Hilton. 2008. Emotional tone and argumentation in risk communication. *Judgment and Decision making* 3, 1 (2008), 100.

[21] Marcin Koszowy, Katarzyna Budzynska, Martín Pereira-Fariña, and Rory Duthie. 2022. From Theory of Rhetoric to the Practice of Language Use: The Case of Appeals to Ethos Elements. *Argumentation* (2022), 1–27.

[22] Klaus Krippendorff, Yann Mathet, Stéphane Bouvry, and Antoine Widlöcher. 2016. On the reliability of unitizing textual continua: Further developments. *Quality & Quantity* 50, 6 (2016), 2347–2364.

[23] Jinfen Li and Lu Xiao. 2020. Emotions in online debates: Tales from 4Forums and ConvinceMe. *Proc. of the Association for IST* 57, 1 (2020), e255.

[24] Marco Lippi and Paolo Torroni. 2016. Argument Mining from Speech: Detecting Claims in Political Debates. *AAAI* 30, 1 (Mar. 2016).

[25] Marco Lippi and Paolo Torroni. 2016. Argument Mining from Speech: Detecting Claims in Political Debates.. In *AAAI*, Vol. 16. 2979–2985.

[26] Stefano Menini, Federico Nanni, Simone Paolo Ponzetto, and Sara Tonelli. 2017. Topic-Based Agreement and Disagreement in US Electoral Manifestos. In *EMNLP*. 2938–2944.

[27] Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 Task 1: Affect in Tweets. In *Proc. of The 12th Int. Workshop on Semantic Evaluation*. 1–17.

[28] Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational intelligence* 29, 3 (2013), 436–465.

[29] Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. 2010. Recognition of Affect, Judgment, and Appreciation in Text. In *Coling*. 806–814.

[30] ALENA NEVIAROUSKAYA, HELMUT PRENDINGER, and MITSURU ISHIZUKA. 2011. Affect Analysis Model: novel rule-based approach to affect sensing from text. *Natural Language Engineering* 17, 1 (2011), 95–135.

[31] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. 1532–1543.

[32] Isaac Persing and Vincent Ng. 2017. Lightly-supervised modeling of argument persuasiveness. In *Proc. of the Eighth Int. Joint Conference on Natural Language Processing*. 594–604.

[33] Peter Potash, Adam Ferguson, and Timothy J. Hazen. 2019. Ranking Passages for Argument Convincingness. In *Proc. of the 6th Workshop on Argument Mining*. 146–155.

[34] Prakash Poudyal, Jaromir Savelka, Aagje Ieven, Marie Francine Moens, Teresa Goncalves, and Paulo Quaresma. 2020. ECHR: Legal Corpus for Argument Mining. In *Proc. of the 7th Workshop on Argument Mining*. 67–75.

[35] C. Rapp. 2002. *Rhetorik*. Number Bd. 1 in Aristoteles Werke in deutscher Übersetzung.

[36] Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and Clustering of Arguments with Contextualized Word Embeddings. In *ACL*. 567–578.

[37] K. Scherer and H. Wallbott. 1994. "Evidence for universality and cultural variation of differential emotion response patterning": Correction. *Journal of Personality and Social Psychology* 67 (1994), 55–55.

[38] Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, et al. 2021. An autonomous debating system. *Nature* 591, 7850 (2021), 379–384.

[39] Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. 2018. Argumentext: Searching for arguments in heterogeneous sources. In *NAACL*. 21–25.

[40] Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic Argument Mining from Heterogeneous Sources. In *Proc. of the 2018 Conf. on EMNLP*. 3664–3674.

[41] Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 Task 14: Affective Text. In *Proc. of the 4th Int. Workshop on Semantic Evaluations* (Prague, Czech Republic) *(SemEval '07)*. 70–74.

[42] Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument Mining: Extracting Arguments from Online Dialogue. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 217–226.

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

[44] S Toulmin. 2003. The Uses of Argument Cambridge University Press.

[45] Dietrich Trautmann, Johannes Daxenberger, Christian Stab, Hinrich Schütze, and Iryna Gurevych. 2020. Fine-Grained Argument Unit Recognition and Classification. In *AAAI*.

[46] Frans H Van Eemeren and Rob Grootendorst. 1987. Fallacies in pragma-dialectical perspective. *Argumentation* 1, 3 (1987), 283–301.

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[48] S Villata et al. 2020. Using Argument Mining for Legal Text Summarization. In *Legal Knowledge and Information Systems: JURIX 2020: The Thirty-third Annual Conf., Brno, Czech Republic, December 9-11, 2020*, Vol. 334. 184.

[49] Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *COLING*. 1680–1691.

[50] Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational Argumentation Quality Assessment in Natural Language. In *EACL*. 176–187.

[51] Henning Wachsmuth, Martin Potthast, Khalid Al Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017. Building an argument search engine for the web. In *Proc. of the 4th Workshop on Argument Mining*. 49–59.

[52] Vern Walker, Karina Vazirova, and Cass Sanford. 2014. Annotating Patterns of Reasoning about Medical Theories of Causation in Vaccine Cases: Toward a Type System for Arguments. In *Proc. of the First Workshop on Argumentation Mining*. 1–10.

[53] Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation schemes*.

[54] Adam Z. Wyner, Raquel Mochales Palau, Marie-Francine Moens, and David Milward. 2010. Approaches to Text Mining Arguments from Legal Cases. In *Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language*, Vol. 6036. 60–79.

A COMPUTING & SOFTWARE INFRASTRUCTURE

All experiments were conducted on a Ubuntu 20.04 system with an AMD Ryzen Processor with 32 CPU-Cores and 126 GB memory. We further used Python 3.7, PyTorch 1.4, and the Huggingface-Transformer library (4.15.0). For the experiments in Chapter 3, we used four NVIDIA RTX 2080 TI GPU with 11 GB memory. The models in Chapter 4 and 5 were trained on a single NVIDIA Tesla V100. The default parameters from the Huggingface-Transformer library [1] were used for all hyperparameters not specified in the following sections.

B GENERALIZATION ACROSS ARGUMENT QUALITY CORPORA

In Section 3, we trained bert-base-uncased models with a batch size of 64. The learning rate was set to $9.1 \cdot 10^{-6}$. A weight decay of 0.31 was used. We calculated the 95th percentile based on the four AQ validation sets and truncated longer sentences to that length. We used a dropout rate of 0.1 for the dropout layer in the **AId → Regression Tasks** setting. The losses in the multi-dataset setting were equally weighted for each of the four datasets. We used early stopping on the validation MSE loss, with a patience value of five epochs, as a regularization technique to avoid over-fitting.

C ZERO-SHOT-LEARNING IN ARGUMENT MINING

For Section 4, we trained bert-large-uncased architectures with a batch size of 64. The learning rate was set to $1 \cdot 10^{-5}$, and for the first 0.1 epochs, a warm-up period is used. We opt for evaluations every 0.1 epochs in our training configuration, resulting in 10 evaluations per epoch. Our train/validation/test split is based on a reasonably standard 70%/10%/20% split. Furthermore, we calculate the 99th percentile of the max length of all sentences inside the validation split, and truncate them to that length. This further decreases the required learning time, due to a reduced input dimension without losing significant information. The losses in the multi-dataset and multi-task setting were equally weighted for each of the three argument mining datasets. Finally, to further reduce variance in training, we use three seeds for our experiments and calculate the mean and standard deviation for all of our results.

D EMOTION DETECTION

For Section 5, we trained bert-base-cased architectures with a batch size of 32. The learning rate was set to $5 \cdot 10^{-5}$ A weight decay of 0.1 was used. We opt for evaluations every 0.25 epochs in our training configuration, resulting in 4 evaluations per epoch. The annotators used the Inception Annotation Framework [2] for the labeling of the arguments. The annotated dataset is split into train/validation/test (60%/10%/30%). Furthermore, we calculate the 99.5th percentile of the max length of all sentences in the validation split, and truncate all sentences to that length.

[1] https://huggingface.co/docs/transformers/master/en/main_classes/trainer#transformers.TrainingArguments
[2] https://inception-project.github.io/releases/22.3/docs/user-guide.html

8 Conclusion

In this thesis, we presented multiple advances in the area of Argument Mining (AM), in particular in the tasks of Argument Identification (AId), Argument Retrieval (AR) and Argument Quality (AQ). We summarize our primary contributions and discuss promising future research directions in the following.

- In Chapter 2, we proposed several methods of using external context information from pre-trained language models and Knowledge Graph embeddings for the task of AId on heterogeneous text domains. We show that external context information is not limited to the argument itself but can further be used for the semantic contextualization of the topic. Our approach allows the consideration of *in- and cross-document* context. Our evaluation shows that external context information provides a drastic performance boost on AId compared to previous state-of-the-art approaches. As a future research direction, we think that other external sources are the main component for a better contextualization of the arguments and the relations in which they exist. Instead of using Knowledge Graph embeddings only to represent the topic, it would be possible also to use it on the argument side. One could investigate a combination of fact-based Knowledge Graphs such as DBPedia[91] or Wikidata [184] with Knowledge Graphs like WordNet [114] and FrameNet [5] that focuses more on the semantical and lexical features. Other pretraining tasks such as *natural language inference* (see Section 1.3.1), which are more similar to the task of AId than masked language modeling, could be beneficial. Lastly, a combination of Knowledge Graph embeddings with pretrained language models could further enhance the performance of AId architectures.

- Chapter 3 introduced a new argument-annotated corpus of peer-reviews from computer science conferences. We show that domain adaptation from another AM dataset is only possible to a limited extent in scientific reviews. An extensive evaluation shows that finetuned models on our corpus can identify arguments in peer-reviews and that the found arguments are essential for the paper acceptance decision. We consider realizing other tasks and architectures on our corpora in future work. Instead of dividing the process of argument detection and relevance detection of the arguments into two steps, an end-to-end model with an adopted objective can learn *both* tasks simultaneously. This would allow the model to learn the role of individual arguments in the decision-making process and estimate its contribution to the acceptance decision. Furthermore, this would allow the ranking of the arguments in the peer-reviews regarding their impact. Especially for the editorial team in the meta-review process, another helpful task would be to identify similar arguments in different peer-reviews of the same paper.

- In Chapter 4, we summarize and compare our work on *AId* in the context of a relational setting. In future work, we could extend the work by broadening the perspective on the generalization capability of AId methods across corpora. To the best of our knowledge, the transfer of trained AId methods between multiple AId corpora has not been studied yet.

- In Chapter 5, we address the AR task. We present a multi-step argument-retrieval system (ARS) that returns relevant and semantically diverse arguments for a given user query in the form of a claim. Our ARS is broadly applicable, it does not rely on an explicit mapping between claims and premises. A further improvement regarding the performance of the ARS could be achieved by a filter-refinement step in front of the relevance filter. Therefore a new query processing would not require the relevance calculation of all stored arguments but instead would be limited to the filtered arguments from *relevant topics*. Additional criteria, besides relevance and diversity, could be added, such as the *stance* of the premises (either supporting or opposing), for a balanced overview. A ranking regarding the AQ could further increase the usefulness of the ARS.

- Chapter 6 addresses the problem of expensive and time-consuming data labeling in argument strength corpora. We present an active-learning approach with uncertainty-based selection strategies. Our work highlights the difficulties of sample-efficient learning in areas that suffer from a rather low agreement between annotators and therefore contain noise in the annotations. We plan to incorporate the agreement scores from single data points in the uncertainty-based selection strategies in future work. This has the effect that the selection strategy uses the most uncertain samples for training *and* samples that fulfill the additional requirement of being a "reliable annotated".

- Whereas previous work in AQ focused on isolated datasets and neglected the interactions with other AM tasks, our work in Chapter 7 fills in this gap. It provides rich results from thorough empirical evaluations. We validate requirements for the generalization across AQ corpora. We further investigate the relations between AQ, AId, and Evidence Detection (ED). Our results show that a zero-shot transfer is to some degree possible. AQ does so far gain less from other tasks than vice-versa. Therefore, an open question is how to achieve a better transfer towards AQ. The three tasks studied in our work can be extended towards other tasks, such as *Stance Detection* [88]. Furthermore, we present a novel AQ corpus annotated with *emotionality* labels. We show that emotions can be detected in argumentation and that emotions play a minor role for AQ than is often assumed. Apart from the well-studied *logos* and the *pathos* dimension investigated in this work, the third remaining dimension, *ethos*, has not received sufficient attention so far. Smaller corpora in this area [40, 86] present a worthwhile objective with AQ.

Since machine learning driven AM is a rather new and unexplored area with numerous applications, there is an excellent potential for future research. In general, we envision

research along the following main axes for the future:

- Research in AM focuses primarily on the *logos* mode of persuasion. Nevertheless, as recent studies showed, the mode of *pathos* represented by emotions is essential in the persuasion process [11, 10, 30, 95, 73]. Cabrio et al. [20] created a corpus of more than 500 arguments annotated by a set of discrete emotions from facial expressions (i.e., happiness, anger, fear, sadness, disgust, and surprise). *Ethos*, which is the last mode of persuasion, is underrepresented in computational AM. Koszowy et al. [86] created a corpus of arguments containing ethos elements such as *practical wisdom, moral virtue*, and *goodwill* which can either be supported or attacked. If these ethotic appeals can be automatically observed by computational applications is still an open question.

- A different axis for future research is the area of *representation learning architectures*. Recent work showed a performance increase in most AM tasks with pretrained language models, compared to previous recurrent architectures, such as the LSTM [74]. Especially on tasks such as AId [49, 148, 178] or AQ estimation [64, 175], contextual embeddings from neural models, such as BERT [33], could provide a more reliable classification compared to previous state-of-the-art architectures. However, most of these works focused on a single AM corpus. A survey with different variations of language models and recurrent architectures could give a further understanding of how the AM tasks can benefit from pretraining tasks.

- Currently, there exist a plethora of small AM corpora based on different text domains and diverse argumentation schemas. However, most of the time, these AM tasks and corpora are studied in *isolation*. Only a few works investigate argument components such as the concept of a "claim" [31] across different corpora. This research direction is critical to learning about the consequences of different conceptualizations of AM theory, especially for practical applications. Furthermore, more research on the connections between the individual AM tasks, such as multi-task experiments, generalization experiments, and multi-objective training, is required to understand how the individual tasks can benefit from each other.

Summarizing, we made contributions towards the reliability of AId (Chapter 2), the expansion of AM domains (Chapter 3), a summarization of relational coarse- and fine-grained AId (Chapter 4), broadly applicable AR (Chapter 5), label efficient AQ estimation (Chapter 6) and contributed towards a more holistic view in AQ (Chapter 7). Our publications, published peer-review corpus, and public codebases ensure that our research directions will be further explored in the future.

Acronyms

AId Argument Identification. 1–3, 6, 12, 13, 16, 18, 19, 21–24, 29, 30, 95–97

AL Active Learning. 30

AM Argument Mining. 1–4, 6, 11–13, 17, 19, 21–23, 26–30, 95–97

AQ Argument Quality. 2, 3, 6, 12, 16, 19, 26, 27, 30, 95–97

AR Argument Retrieval. 2, 3, 16, 19, 24, 25, 29, 30, 95–97

ARS Argument Retrieval System. 24, 25, 30

BERT Bidirectional Encoder Representations from Transformers. 13, 17, 18, 27

BiLSTM bidirectional Long Short-Term Memory. 27

CBOW Continuous Bag-of-Words. 6, 7

CCE Categorical Cross-Entropy. 6

CLSTM Contextual Long Short-Term Memory. 10, 12, 13, 23

ED Evidence Detection. 30, 96

GloVe Global Vectors. 6, 7, 9, 27

LM language modeling. 17, 18

LSA latent semantic analysis. 7

LSTM Long Short-Term Memory. 10–13, 23

ML Machine Learning. 1–3, 16, 19, 29, 30

MLM masked-language modeling. 2, 16

MLP Multilayer Perceptron. 5, 6, 10, 27

MSE Mean squared error. 6

NLP Natural Language Processing. 1–4, 13, 16–18, 21, 23, 28, 29

NSP next sentence prediction. 2, 16, 18

ReLU rectified linear unit. 5, 10

RNN Recurrent Neural Network. 3, 6, 10, 11, 13

SNLI Stanford Natural Language Inference. 17, 18

SVM Support-vector machines. 3, 6, 26, 27

TF-IDF term frequency-inverse document frequency. 4

Bibliography

[1] Abien Fred Agarap. „Deep learning using rectified linear units (relu).“ In: *arXiv preprint arXiv:1803.08375* (2018).

[2] Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. „Where is your evidence: improving fact-checking by justification modeling.“ In: *Proceedings of the first workshop on fact extraction and verification (FEVER)*. 2018, pp. 85–90.

[3] Rawia Awadallah, Maya Ramanath, and Gerhard Weikum. „Harmony and Dissonance: Organizing the People's Voices on Political Controversies.“ In: *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*. WSDM '12. Seattle, Washington, USA: Association for Computing Machinery, 2012, pp. 523–532. ISBN: 9781450307475. DOI: 10.1145/2124295.2124359. URL: https://doi.org/10.1145/2124295.2124359.

[4] Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. „Neural machine translation by jointly learning to align and translate.“ In: *3rd International Conference on Learning Representations, ICLR 2015*. 2015.

[5] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. „The Berkeley FrameNet Project.“ In: *Proceedings of the 17th International Conference on Computational Linguistics - Volume 1*. COLING '98. Montreal, Quebec, Canada: Association for Computational Linguistics, 1998, pp. 86–90. DOI: 10.3115/980451.980860. URL: https://doi.org/10.3115/980451.980860.

[6] 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*. Online: Association for Computational Linguistics, July 2020, pp. 4029–4039. DOI: 10.18653/v1/2020.acl-main.371. URL: https://aclanthology.org/2020.acl-main.371.

[7] 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*. 2020, pp. 4029–4039.

[8] Roy Bar-Haim, Yoav Kantor, Lilach Eden, Roni Friedman, Dan Lahav, and Noam Slonim. „Quantitative argument summarization and beyond: Cross-domain key point analysis.“ In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020, pp. 39–49.

[9] Anna Beniermann, Laurens Mecklenburg, and Annette Upmeier zu Belzen. „Reasoning on Controversial Science Issues in Science Education and Science Communication.“ In: *Education Sciences* 11.9 (2021). ISSN: 2227-7102. DOI: 10.3390/educsci11090522. URL: https://www.mdpi.com/2227-7102/11/9/522.

[10] Mohamed S. Benlamine, Serena Villata, Ramla Ghali, Claude Frasson, Fabien Gandon, and Elena Cabrio. „Persuasive Argumentation and Emotions: An Empirical Evaluation with Users.“ In: *Human-Computer Interaction. User Interface Design, Development and Multimodality*. Ed. by Masaaki Kurosu. Cham: Springer International Publishing, 2017, pp. 659–671. ISBN: 978-3-319-58071-5.

[11] Sahbi Benlamine, Maher Chaouachi, Serena Villata, Elena Cabrio, Claude Frasson, and Fabien Gandon. „Emotions in argumentation: an empirical evaluation.“ In: *Twenty-Fourth International Joint Conference on Artificial Intelligence*. 2015.

[12] Floris Bex, John Lawrence, Mark Snaith, and Chris Reed. „Implementing the Argument Web.“ In: *Commun. ACM* 56.10 (Oct. 2013), pp. 66–73. ISSN: 0001-0782. DOI: 10.1145/2500891. URL: https://doi.org/10.1145/2500891.

[13] J. Anthony Blair. „Argumentation as Rational Persuasion.“ In: *Argumentation* 26.1 (2012), pp. 71–81. DOI: 10.1007/s10503-011-9235-6.

[14] Alexander Bondarenko, Maik Fröbe, Meriem Beloucif, Lukas Gienapp, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, et al. „Overview of Touché 2020: Argument Retrieval.“ In: *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer. 2020, pp. 384–395.

[15] Alexander Bondarenko, Lukas Gienapp, Maik Fröbe, Meriem Beloucif, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, et al. „Overview of Touché 2021: argument retrieval.“ In: *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer. 2021, pp. 450–467.

[16] Tom Bosc, Elena Cabrio, and Serena Villata. „Tweeties Squabbling: Positive and Negative Results in Applying Argument Mining on Social Media.“ In: *Computational Models of Argument: Proc. from the Sixth Int. Conference on Computational Models of Argument (COMMA)* 2016 (2016), pp. 21–32.

[17] Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. „A large annotated corpus for learning natural language inference.“ In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2015, pp. 632–642.

[18] Leo Breiman. „Random forests.“ In: *Machine learning* 45.1 (2001), pp. 5–32.

[19] Maximilian Bundesmann, Lukas Christ, and Matthias Richter. „Creating an Argument Search Engine for Online Debates.“ In: *Conference and Labs of the Evaluation Forum (CLEF) (Working Notes)*. 2020.

[20] Elena Cabrio and Serena Villata. „The SEEMPAD Dataset for Emphatic and Persuasive Argumentation.“ In: *Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it*. Vol. 10. 2018, p. 12.

[21] Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. „Give Me More Feedback: Annotating Argument Persuasiveness and Related Attributes in Student Essays.“ In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 621–631. DOI: 10.18653/v1/P18-1058. URL: https://aclanthology.org/P18-1058.

[22] Lucas Carstens and Francesca Toni. „Towards relation based argumentation mining.“ In: *Proceedings of the 2nd Workshop on Argumentation Mining*. 2015, pp. 29–34.

[23] Lucas Carstens, Francesca Toni, and Valentinos Evripidou. „Argument mining and social debates.“ In: *Computational Models of Argument*. IOS Press, 2014, pp. 451–452.

[24] Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathy McKeown, and Alyssa Hwang. „AMPERSAND: Argument Mining for PERSuAsive oNline Discussions.“ 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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 2933–2943. DOI: 10.18653/v1/D19-1291. URL: https://aclanthology.org/D19-1291.

[25] Artem Chernodub, Oleksiy Oliynyk, Philipp Heidenreich, Alexander Bondarenko, Matthias Hagen, Chris Biemann, and Alexander Panchenko. „Targer: Neural argument mining at your fingertips.“ In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 2019, pp. 195–200.

[26] Yoonjung Choi, Yuchul Jung, and Sung-Hyon Myaeng. „Identifying Controversial Issues and Their Sub-topics in News Articles.“ In: *Intelligence and Security Informatics*. Ed. by Hsinchun Chen, Michael Chau, Shu-hsing Li, Shalini Urs, Srinath Srinivasa, and G. Alan Wang. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 140–153. ISBN: 978-3-642-13601-6.

[27] Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. „Novelty and Diversity in Information Retrieval Evaluation.“ In: *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '08. Singapore, Singapore: Association for Computing Machinery, 2008, pp. 659–666. ISBN: 9781605581644. DOI: 10.1145/1390334.1390446. URL: https://doi.org/10.1145/1390334.1390446.

[28] Corinna Cortes and Vladimir Vapnik. „Support vector machine.“ In: *Machine learning* 20.3 (1995), pp. 273–297.

[29] George Cybenko. „Approximation by superpositions of a sigmoidal function.“ In: *Mathematics of control, signals and systems* 2.4 (1989), pp. 303–314.

[30] Francesca D'Errico, Marinella Paciello, and Matteo Amadei. „Behind Our Words: Psychological Paths Underlying the Un/Supportive Stance Toward Immigrants in Social Media.“ In: *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. 2018, pp. 649–656. DOI: 10.1109/DSAA.2018.00084.

[31] 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*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 2055–2066. DOI: 10.18653/v1/D17-1218. URL: https://aclanthology.org/D17-1218.

[32] Johannes Daxenberger, Benjamin Schiller, Chris Stahlhut, Erik Kaiser, and Iryna Gurevych. „Argumentext: argument classification and clustering in a generalized search scenario.“ In: *Datenbank-Spektrum* 20.2 (2020), pp. 115–121.

[33] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.“ In: *NAACL-HLT (1)*. 2019.

[34] Susan T Dumais, George W Furnas, Thomas K Landauer, Scott Deerwester, and Richard Harshman. „Using latent semantic analysis to improve access to textual information.“ In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1988, pp. 281–285.

[35] Lorik Dumani, Patrick J Neumann, and Ralf Schenkel. „A Framework for Argument Retrieval.“ In: *European Conference on Information Retrieval*. Springer. 2020, pp. 431–445.

[36] Lorik Dumani and Ralf Schenkel. „A systematic comparison of methods for finding good premises for claims.“ In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2019, pp. 957–960.

[37] Lorik Dumani and Ralf Schenkel. „Quality-aware ranking of arguments.“ In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2020, pp. 335–344.

[38] Mihai Dusmanu, Elena Cabrio, and Serena Villata. „Argument mining on Twitter: Arguments, facts and sources.“ In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017, pp. 2317–2322.

[39] Rory Duthie, Katarzyna Budzynska, and Chris Reed. „Mining Ethos in Political Debate.“ English. In: *Computational Models of Argument*. Ed. by Pietro Baroni, Thomas F. Gordon, Tatjana Scheffler, and Manfred Stede. Vol. 287. Frontiers in Artificial Intelligence and Applications. This research was supported in part by EPSRC in the UK under grant EP/M506497/1 and in part by the Polish National

Science Centre under grant 2015/18/M/HS1/00620. Netherlands: IOS Press, 2016, pp. 299–310. ISBN: 9781614996859. DOI: 10.3233/978-1-61499-686-6-299.

[40] Rory Duthie, Katarzyna Budzynska, and Chris Reed. „Mining ethos in political debate.“ In: *Computational Models of Argument: Proc. from the Sixth Int. Conference on Computational Models of Argument (COMMA)*. 2016, pp. 299–310.

[41] Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. „Neural End-to-End Learning for Computational Argumentation Mining.“ In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2017, pp. 11–22.

[42] Liat Ein-Dor, Eyal Shnarch, Lena Dankin, Alon Halfon, Benjamin Sznajder, Ariel Gera, Carlos Alzate, Martin Gleize, Leshem Choshen, Yufang Hou, et al. „Corpus wide argument mining—a working solution.“ In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05. 2020, pp. 7683–7691.

[43] Jeffrey L Elman. „Finding structure in time.“ In: *Cognitive science* 14.2 (1990), pp. 179–211.

[44] Adam Robert Faulkner. *Automated classification of argument stance in student essays: A linguistically motivated approach with an application for supporting argument summarization*. City University of New York, 2014.

[45] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. „Placing search in context: The concept revisited.“ In: *Proceedings of the 10th international conference on World Wide Web*. 2001, pp. 406–414.

[46] J. R. Firth. „THE TECHNIQUE OF SEMANTICS.“ In: *Transactions of the Philological Society* 34.1 (1935), pp. 36–73. DOI: https://doi.org/10.1111/j.1467-968X.1935.tb01254.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-968X.1935.tb01254.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-968X.1935.tb01254.x.

[47] **Michael Fromm**, Max Berrendorf, Johanna Reiml, Isabelle Mayerhofer, Siddharth Bhargava, Evgeniy Faerman, and Thomas Seidl. „Towards a Holistic View on Argument Quality Prediction.“ In: Peer-Reviewing Phase.

[48] **Michael Fromm**, Evgeniy Faerman, Max Berrendorf, Siddharth Bhargava, Ruoxia Qi, Yao Zhang, Lukas Dennert, Sophia Selle, Yang Mao, and Thomas Seidl. „Argument Mining Driven Analysis of Peer-Reviews.“ In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35 (May 2021), pp. 4758–4766. URL: https://ojs.aaai.org/index.php/AAAI/article/view/16607.

[49] **Michael Fromm**, Evgeniy Faerman, and Thomas Seidl. „TACAM: Topic And Context Aware Argument Mining.“ In: *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE. 2019, pp. 99–106.

[50] **Michael Fromm***, Max Berrendorf*, Sandra Obermeier, Thomas Seidl, and Evgeniy Faerman. „Diversity Aware Relevance Learning for Argument Search.“ In: *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part II*. Vol. 12657. Lecture Notes in Computer Science. * equal contribution. Springer, 2021, pp. 264–271. DOI: 10.1007/978-3-030-72240-1_24. URL: https://doi.org/10.1007/978-3-030-72240-1_24.

[51] Yarin Gal and Zoubin Ghahramani. „Dropout as a bayesian approximation: Representing model uncertainty in deep learning.“ In: *international conference on machine learning*. PMLR. 2016, pp. 1050–1059.

[52] Andrea Galassi, Marco Lippi, and Paolo Torroni. „Argumentative link prediction using residual networks and multi-objective learning.“ In: *Proceedings of the 5th Workshop on Argument Mining*. 2018, pp. 1–10.

[53] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. „Convolutional sequence to sequence learning.“ In: *International Conference on Machine Learning*. PMLR. 2017, pp. 1243–1252.

[54] Tim van Gelder. „The rationale for Rationale™.“ In: *Law, Probability and Risk* 6.1-4 (Oct. 2007), pp. 23–42. ISSN: 1470-8396. DOI: 10.1093/lpr/mgm032. eprint: https://academic.oup.com/lpr/article-pdf/6/1-4/23/2852954/mgm032.pdf. URL: https://doi.org/10.1093/lpr/mgm032.

[55] Lise Getoor and Christopher P Diehl. „Link mining: a survey.“ In: *Acm Sigkdd Explorations Newsletter* 7.2 (2005), pp. 3–12.

[56] Shalini Ghosh, Oriol Vinyals, Brian Strope, Scott Roy, Tom Dean, and Larry Heck. „Contextual lstm (clstm) models for large scale nlp tasks.“ In: *arXiv preprint arXiv:1602.06291* (2016).

[57] Matthew Gifford. „Lexridelaw: an argument based legal search engine.“ In: *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*. 2017, pp. 271–272.

[58] 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 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 967–976. DOI: 10.18653/v1/P19-1093. URL: https://www.aclweb.org/anthology/P19-1093.

[59] Amir Globerson, Gal Chechik, Fernando Pereira, and Naftali Tishby. „Euclidean embedding of co-occurrence data.“ In: *Advances in neural information processing systems (NeurIPS)* 17 (2004).

[60] Xavier Glorot and Yoshua Bengio. „Understanding the difficulty of training deep feedforward neural networks.“ In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2010, pp. 249–256.

[61] Thomas F. Gordon, Henry Prakken, and Douglas Walton. „The Carneades Model of Argument and Burden of Proof.“ In: *Artif. Intell.* 171.10–15 (July 2007), pp. 875–896. ISSN: 0004-3702. DOI: 10.1016/j.artint.2007.04.010. URL: https://doi.org/10.1016/j.artint.2007.04.010.

[62] Theodosis Goudas, Christos Louizos, Georgios Petasis, and Vangelis Karkaletsis. „Argument extraction from news, blogs, and social media.“ In: *Hellenic Conference on Artificial Intelligence*. Springer. 2014, pp. 287–299.

[63] Klaus Greff, Rupesh K Srivastava, Jan Koutnik, Bas R Steunebrink, and Jürgen Schmidhuber. „LSTM: A search space odyssey.“ In: *IEEE transactions on neural networks and learning systems* 28.10 (2016), pp. 2222–2232.

[64] Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. „A large-scale dataset for argument quality ranking: Construction and analysis.“ In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05. 2020, pp. 7805–7813.

[65] Chinnappa Guggilla, Tristan Miller, and Iryna Gurevych. „CNN-and LSTM-based claim classification in online user comments.“ In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 2016, pp. 2740–2751.

[66] Kristina Gulordava, Piotr Bojanowski, Édouard Grave, Tal Linzen, and Marco Baroni. „Colorless Green Recurrent Networks Dream Hierarchically.“ In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 2018, pp. 1195–1205.

[67] Ivan Habernal and Iryna Gurevych. „Argumentation mining in user-generated web discourse.“ In: *Computational Linguistics* 43.1 (2017), pp. 125–179.

[68] 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*. Austin, Texas: Association for Computational Linguistics, 2016, pp. 1214–1223. DOI: 10.18653/v1/D16-1129. URL: https://www.aclweb.org/anthology/D16-1129.

[69] 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*. 2016, pp. 1214–1223.

[70] Ivan Habernal and Iryna Gurevych. „Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm.“ In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2016, pp. 1589–1599.

[71] Shohreh Haddadan, Elena Cabrio, and Serena Villata. „Yes, we can! Mining Arguments in 50 Years of US Presidential Campaign Debates." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 4684–4690. DOI: 10.18653/v1/P19-1463. URL: https://aclanthology.org/P19-1463.

[72] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. „Delving deep into rectifiers: Surpassing human-level performance on imagenet classification." In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1026–1034.

[73] Denis Hilton. „Emotional tone and argumentation in risk communication." In: *Judgment and Decision making* 3.1 (2008), p. 100.

[74] Sepp Hochreiter and Jürgen Schmidhuber. „Long short-term memory." In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[75] Alexander Hogenboom, Frederik Hogenboom, Uzay Kaymak, Paul Wouters, and Franciska De Jong. „Mining economic sentiment using argumentation structures." In: *International Conference on Conceptual Modeling*. Springer. 2010, pp. 200–209.

[76] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. „Multilayer feedforward networks are universal approximators." In: *Neural networks* 2.5 (1989), pp. 359–366.

[77] Yufang Hou and Charles Jochim. „Argument relation classification using a joint inference model." In: *Proceedings of the 4th Workshop on Argument Mining*. 2017, pp. 60–66.

[78] Jeremy Howard and Sebastian Ruder. „Universal Language Model Fine-tuning for Text Classification." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018, pp. 328–339.

[79] Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. „Improving word representations via global context and multiple word prototypes." In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2012, pp. 873–882.

[80] Amalia Huwaidah, Said Al Faraby, et al. „Argument Identification in Indonesian Tweets on the Issue of Moving the Indonesian Capital." In: *Procedia Computer Science* 179 (2021), pp. 407–415.

[81] Neslihan Iskender, Robin Schaefer, Tim Polzehl, and Sebastian Möller. „Argument Mining in Tweets: Comparing Crowd and Expert Annotations for Automated Claim and Evidence Detection." In: *International Conference on Applications of Natural Language to Information Systems*. Springer. 2021, pp. 275–288.

[82] George H. John and Pat Langley. „Estimating Continuous Distributions in Bayesian Classifiers." In: *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. UAI'95. Montréal, Qué, Canada: Morgan Kaufmann Publishers Inc., 1995, pp. 338–345. ISBN: 1558603859.

[83] Nataliia Kees, **Michael Fromm**, Evgeniy Faerman, and Thomas Seidl. „Active Learning for Argument Strength Estimation." In: *Proceedings of the Second Workshop on Insights from Negative Results in NLP*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 144–150. URL: https://aclanthology.org/2021.insights-1.20.

[84] Victor Alvin Ketcham. *The theory and practice of argumentation and debate,* eng. New York: The Macmillan company, 1925.

[85] Aniket Kittur, Bongwon Suh, Bryan Pendleton, and Ed Chi. „He Says, She Says: Conflict and Coordination in Wikipedia." In: Apr. 2007. DOI: 10.1145/1240624.1240698.

[86] Marcin Koszowy, Katarzyna Budzynska, Martin Pereira-Fariña, and Rory Duthie. „From Theory of Rhetoric to the Practice of Language Use: The Case of Appeals to Ethos Elements." In: *Argumentation* (2022), pp. 1–27.

[87] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. „Imagenet classification with deep convolutional neural networks." In: *Advances in neural information processing systems* 25 (2012), pp. 1097–1105.

[88] Dilek Küçük and Fazli Can. „Stance detection: A survey." In: *ACM Computing Surveys (CSUR)* 53.1 (2020), pp. 1–37.

[89] Saskia Le Cessie and Johannes C Van Houwelingen. „Ridge estimators in logistic regression." In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 41.1 (1992), pp. 191–201.

[90] Rémi Lebret and Ronan Collobert. „Word Embeddings through Hellinger PCA." In: *European ACL, (EACL)*. 2014.

[91] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. „DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia." In: *Semantic Web Journal* 6.2 (2015), pp. 167–195. URL: http://jens-lehmann.org/files/2015/swj_dbpedia.pdf.

[92] Omer Levy and Yoav Goldberg. „Linguistic regularities in sparse and explicit word representations." In: *Proceedings of the eighteenth conference on computational natural language learning*. 2014, pp. 171–180.

[93] Omer Levy and Yoav Goldberg. „Neural word embedding as implicit matrix factorization." In: *Advances in neural information processing systems* 27 (2014), pp. 2177–2185.

[94] Ran Levy, Ben Bogin, 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*. 2018, pp. 2066–2081.

[95] Jinfen Li and Lu Xiao. „Emotions in online debates: Tales from 4Forums and ConvinceMe." In: *Proceedings of the Association for Information Science and Technology* 57.1 (2020), e255.

[96] Yitan Li, Linli Xu, Fei Tian, Liang Jiang, Xiaowei Zhong, and Enhong Chen. „Word embedding revisited: A new representation learning and explicit matrix factorization perspective.“ In: *Twenty-Fourth International Joint Conference on Artificial Intelligence*. 2015.

[97] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. „Learning entity and relation embeddings for knowledge graph completion.“ In: *Twenty-ninth AAAI conference on artificial intelligence*. 2015.

[98] Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. „Assessing the ability of LSTMs to learn syntax-sensitive dependencies.“ In: *Transactions of the Association for Computational Linguistics* 4 (2016), pp. 521–535.

[99] Marco Lippi and Paolo Torroni. „Argument Mining from Speech: Detecting Claims in Political Debates.“ In: *Proceedings of the AAAI Conference on Artificial Intelligence* 30.1 (Mar. 2016). URL: `https://ojs.aaai.org/index.php/AAAI/article/view/10384`.

[100] Marco Lippi and Paolo Torroni. „Argument mining: A machine learning perspective.“ In: *International Workshop on Theory and Applications of Formal Argumentation*. Springer. 2015, pp. 163–176.

[101] Marco Lippi and Paolo Torroni. „MARGOT: A web server for argumentation mining.“ In: *Expert Systems with Applications* 65 (2016), pp. 292–303.

[102] Bing Liu et al. „Sentiment analysis and subjectivity.“ In: *Handbook of natural language processing* 2.2010 (2010), pp. 627–666.

[103] Minh-Thang Luong, Richard Socher, and Christopher D Manning. „Better word representations with recursive neural networks for morphology.“ In: *Proceedings of the seventeenth conference on computational natural language learning*. 2013, pp. 104–113.

[104] Elias J MacEwan. *The essentials of argumentation*. DC Heath & Company, 1898.

[105] Tobias Mayer, Elena Cabrio, Marco Lippi, Paolo Torroni, and Serena Villata. „Argument Mining on Clinical Trials.“ In: *Computational Models of Argument: Proc. from the 7th Int. Conference on Computational Models of Argument (COMMA)*. 2018, pp. 137–148.

[106] Tobias Mayer, Elena Cabrio, and Serena Villata. „Transformer-based argument mining for healthcare applications.“ In: *ECAI 2020*. IOS Press, 2020, pp. 2108–2115.

[107] Hugo Mercier and Dan Sperber. „Why do humans reason? Arguments for an argumentative theory.“ In: *Behavioral and brain sciences* 34.2 (2011), pp. 57–74.

[108] Stephen Merity, Tara Murphy, and James R Curran. „Accurate argumentative zoning with maximum entropy models.“ In: *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries (NLPIR4DL)*. 2009, pp. 19–26.

[109] Andrei Mikheev, Claire Grover, and Marc Moens. „Description of the LTG System Used for MUC-7.“ In: *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*. 1998. URL: https://aclanthology.org/M98-1021.

[110] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. „Efficient estimation of word representations in vector space.“ In: *arXiv preprint arXiv:1301.3781* (2013).

[111] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. „Distributed Representations of Words and Phrases and Their Compositionality.“ In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'13. Lake Tahoe, Nevada: Curran Associates Inc., 2013, pp. 3111–3119.

[112] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. „Distributed representations of words and phrases and their compositionality.“ In: *Advances in neural information processing systems*. 2013, pp. 3111–3119.

[113] George A Miller and Walter G Charles. „Contextual correlates of semantic similarity.“ In: *Language and cognitive processes* 6.1 (1991), pp. 1–28.

[114] George A. Miller. „WordNet: A Lexical Database for English.“ In: *Commun. ACM* 38.11 (Nov. 1995), pp. 39–41. ISSN: 0001-0782. DOI: 10.1145/219717.219748. URL: http://doi.acm.org/10.1145/219717.219748.

[115] Amita Misra, Pranav Anand, Jean E Fox Tree, and Marilyn Walker. „Using Summarization to Discover Argument Facets in Online Idealogical Dialog.“ In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2015, pp. 430–440.

[116] 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*. 2007, pp. 225–230.

[117] Sebastian Möller. „Argument Mining in Tweets: Comparing Crowd and Expert Annotations for Automated Claim and Evidence Detection.“ In: *Natural Language Processing and Information Systems: 26th International Conference on Applications of Natural Language to Information Systems, NLDB 2021, Saarbrücken, Germany, June 23–25, 2021, Proceedings*. Vol. 12801. Springer Nature. 2021, p. 275.

[118] Gaku Morio and Katsuhide Fujita. „End-to-End Argument Mining for Discussion Threads Based on Parallel Constrained Pointer Architecture.“ In: *Proceedings of the 5th Workshop on Argument Mining*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 11–21. DOI: 10.18653/v1/W18-5202. URL: https://aclanthology.org/W18-5202.

[119] Huy Nguyen and Diane Litman. „Argument mining for improving the automated scoring of persuasive essays.“ In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1. 2018.

[120] Huy Nguyen and Diane Litman. „Context-aware argumentative relation mining.“ In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2016, pp. 1127–1137.

[121] Huy Nguyen and Diane Litman. „Improving argument mining in student essays by learning and exploiting argument indicators versus essay topics.“ In: *The Twenty-Ninth International Flairs Conference*. 2016.

[122] D.J. O'Keefe. *Persuasion: Theory and Research*. Current communication. Sage Publications, 2002. ISBN: 9780761925392. URL: https://books.google.de/books?id=e3V6Zen0UGwC.

[123] Nathan Ong, Diane Litman, and Alexandra Brusilovsky. „Ontology-based argument mining and automatic essay scoring.“ In: *Proceedings of the First Workshop on Argumentation Mining*. 2014, pp. 24–28.

[124] Juri Opitz and Anette Frank. „Dissecting Content and Context in Argumentative Relation Analysis.“ In: *Proceedings of the 6th Workshop on Argument Mining*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 25–34. DOI: 10.18653/v1/W19-4503. URL: https://aclanthology.org/W19-4503.

[125] Juri Opitz and Anette Frank. „Dissecting Content and Context in Argumentative Relation Analysis.“ In: *Proceedings of the 6th Workshop on Argument Mining*. 2019, pp. 25–34.

[126] Sebastian Padó, Tae-Gil Noh, Asher Stern, Rui Wang, and Roberto Zanoli. „Design and realization of a modular architecture for textual entailment.“ In: *Natural Language Engineering* 21.2 (2015), pp. 167–200.

[127] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. „On the difficulty of training recurrent neural networks.“ In: *International conference on machine learning*. PMLR. 2013, pp. 1310–1318.

[128] Jeffrey Pennington, Richard Socher, and Christopher D Manning. „Glove: Global vectors for word representation.“ In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543.

[129] 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*. 2016, pp. 1384–1394.

[130] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. „Deep Contextualized Word Representations.“ In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 2227–2237. DOI: 10.18653/v1/N18-1202. URL: https://aclanthology.org/N18-1202.

[131] Scott Piao, Sophia Ananiadou, Yoshimasa Tsuruoka, Yutaka Sasaki, and John McNaught. „Mining opinion polarity relations of citations.“ In: *International Workshop on Computational Semantics (IWCS)*. 2007, pp. 366–371.

[132] Peter Potash, Robin Bhattacharya, and Anna Rumshisky. „Length, Interchangeability, and External Knowledge: Observations from Predicting Argument Convincingness.“ In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, Nov. 2017, pp. 342–351. URL: https://aclanthology.org/I17-1035.

[133] Peter Potash, Adam Ferguson, and Timothy J. Hazen. „Ranking Passages for Argument Convincingness.“ In: *Proceedings of the 6th Workshop on Argument Mining*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 146–155. DOI: 10.18653/v1/W19-4517. URL: https://aclanthology.org/W19-4517.

[134] Peter Potash, Alexey Romanov, and Anna Rumshisky. „Here's My Point: Argumentation Mining with Pointer Networks.“ In: (2016).

[135] 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*. Online: Association for Computational Linguistics, Dec. 2020, pp. 67–75. URL: https://aclanthology.org/2020.argmining-1.8.

[136] J Ross Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.

[137] M Atif Qureshi and Derek Greene. „Eve: explainable vector based embedding technique using wikipedia.“ In: *Journal of Intelligent Information Systems* 53.1 (2019), pp. 137–165.

[138] Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. „Learning to generate reviews and discovering sentiment.“ In: *arXiv preprint arXiv:1704.01444* (2017).

[139] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. „Improving language understanding by generative pre-training.“ In: *Preprint* ().

[140] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. „Exploring the limits of transfer learning with a unified text-to-text transformer.“ In: *J. Mach. Learn. Res.* 21.140 (2020), pp. 1–67.

[141] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. „SQuAD: 100,000+ Questions for Machine Comprehension of Text.“ In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2016, pp. 2383–2392.

[142] C. Rapp. *Rhetorik*. Aristoteles Werke in deutscher Übersetzung Bd. 1. Akademie Verlag, 2002. ISBN: 9783050037011. URL: https://books.google.de/books?id=4wHWvQEACAAJ.

[143] Christof Rapp. „Aristotle's Rhetoric." In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2010. Metaphysics Research Lab, Stanford University, 2010.

[144] Shauli Ravfogel, Yoav Goldberg, and Tal Linzen. „Studying the Inductive Biases of RNNs with Synthetic Variations of Natural Languages." In: *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT*. 2019, pp. 3532–3542.

[145] Chris Reed. „Argument technology for debating with humans." In: *Nature* 591.7850 (2021), pp. 373–374.

[146] Chris Reed. „Preliminary results from an argument corpus." In: *Linguistics in the twenty-first century* (2006), pp. 185–196.

[147] 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*. 2019, pp. 567–578.

[148] 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*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 567–578. DOI: 10.18653/v1/P19-1054. URL: https://aclanthology.org/P19-1054.

[149] Allen Roush and Arvind Balaji. „DebateSum: A large-scale argument mining and summarization dataset." In: *Proceedings of the 7th Workshop on Argument Mining*. Online: Association for Computational Linguistics, Dec. 2020, pp. 1–7. URL: https://aclanthology.org/2020.argmining-1.1.

[150] Glenn Rowe, Fabrizio Macagno, Chris Reed, and Douglas Walton. „Araucaria as a Tool for Diagramming Arguments in Teaching and Studying Philosophy." In: *Teaching Philosophy* 29 (2006), pp. 111–124.

[151] Herbert Rubenstein and John B Goodenough. „Contextual correlates of synonymy." In: *Communications of the ACM* 8.10 (1965), pp. 627–633.

[152] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. „Learning representations by back-propagating errors." In: *nature* 323.6088 (1986), pp. 533–536.

[153] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. „ImageNet Large Scale Visual Recognition Challenge." In: *International Journal of Computer Vision (IJCV)* 115.3 (2015), pp. 211–252. DOI: 10.1007/s11263-015-0816-y.

[154] Patrick Saint-Dizier. „A Two-Level Approach to Generate Synthetic Argumentation Reports." In: *Argument Comput.* 9 (2017), pp. 137–154.

[155] Patrick Saint-Dizier. „Knowledge-driven argument mining based on the qualia structure.“ In: *Argument & Computation* 8.2 (2017), pp. 193–210.

[156] Nattapong Sanchan, Ahmet Aker, and Kalina Bontcheva. „Automatic Summarization of Online Debates.“ In: *Proceedings of the 1st Workshop on Natural Language Processing and Information Retrieval associated with RANLP 2017*. Varna, Bulgaria: INCOMA Inc., Sept. 2017, pp. 19–27. URL: https://doi.org/10.26615/978-954-452-038-0_003.

[157] Erik Tjong Kim Sang and Fien De Meulder. „Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition.“ In: *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*. 2003, pp. 142–147.

[158] Christos Sardianos, Ioannis Manousos Katakis, Georgios Petasis, and Vangelis Karkaletsis. „Argument extraction from news.“ In: *Proceedings of the 2nd Workshop on Argumentation Mining*. 2015, pp. 56–66.

[159] Robin Schaefer and Manfred Stede. „Annotation and detection of arguments in tweets.“ In: *Proceedings of the 7th Workshop on Argument Mining*. 2020, pp. 53–58.

[160] Robin Schaefer and Manfred Stede. „Argument Mining on Twitter: A survey.“ In: *it-Information Technology* 63.1 (2021), pp. 45–58.

[161] Robin Schaefer and Manfred Stede. „Improving implicit stance classification in tweets using word and sentence embeddings.“ In: *Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz)*. Springer. 2019, pp. 299–307.

[162] E. Schiappa and J.P. Nordin. *Argumentation: Keeping Faith with Reason*. Pearson, 2014. ISBN: 9780205327447. URL: https://books.google.de/books?id=e9-9kgEACAAJ.

[163] Claudia Schulz, Steffen Eger, Johannes Daxenberger, Tobias Kahse, and Iryna Gurevych. „Multi-Task Learning for Argumentation Mining in Low-Resource Settings.“ In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 35–41. DOI: 10.18653/v1/N18-2006. URL: https://aclanthology.org/N18-2006.

[164] Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. „CCMatrix: Mining Billions of High-Quality Parallel Sentences on the Web.“ In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 6490–6500. DOI: 10.18653/v1/2021.acl-long.507. URL: https://aclanthology.org/2021.acl-long.507.

[165] Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, et al. „An autonomous debating system." In: *Nature* 591.7850 (2021), pp. 379–384.

[166] 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: demonstrations*. 2018, pp. 21–25.

[167] 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, pp. 46–56.

[168] Christian Stab, Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. „Argumentation Mining in Persuasive Essays and Scientific Articles from the Discourse Structure Perspective." In: *Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing (ArgNLP)*. 2014, pp. 21–25.

[169] Christian Stab, Tristan Miller, and Iryna Gurevych. „Cross-topic Argument Mining from Heterogeneous Sources Using Attention-based Neural Networks." In: *CoRR* abs/1802.05758 (2018). arXiv: 1802.05758. URL: http://arxiv.org/abs/1802.05758.

[170] 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*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 3664–3674. DOI: 10.18653/v1/D18-1402. URL: https://aclanthology.org/D18-1402.

[171] Benno Stein, Yamen Ajjour, Roxanne El Baff, Khalid Al-Khatib, Philipp Cimiano, AG Semantic Computing, and Henning Wachsmuth. „Same Side Stance Classification." In: *Preprint* (2021).

[172] Ola Svenson. „Process descriptions of decision making." In: *Organizational Behavior and Human Performance* 23.1 (1979), pp. 86–112. ISSN: 0030-5073. DOI: https://doi.org/10.1016/0030-5073(79)90048-5. URL: https://www.sciencedirect.com/science/article/pii/0030507379900485.

[173] Reid Swanson, Brian Ecker, and Marilyn Walker. „Argument Mining: Extracting Arguments from Online Dialogue." In: *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Prague, Czech Republic: Association for Computational Linguistics, Sept. 2015, pp. 217–226. DOI: 10.18653/v1/W15-4631. URL: https://aclanthology.org/W15-4631.

[174] Simone Teufel, Advaith Siddharthan, and Colin Batchelor. „Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics." In: *Proceedings of the 2009 conference on empirical methods in natural language processing*. 2009, pp. 1493–1502.

[175] 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.“ 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)*. 2019, pp. 5625–5635.

[176] Lisa Torrey and Jude Shavlik. „Transfer learning.“ In: *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI global, 2010, pp. 242–264.

[177] Stephen E Toulmin. *The uses of argument*. Cambridge university press, 1958.

[178] 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* 34.05 (Apr. 2020), pp. 9048–9056. DOI: 10.1609/aaai.v34i05.6438. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6438.

[179] Dietrich Trautmann, **Michael Fromm**, Volker Tresp, Thomas Seidl, and Hinrich Schütze. „Relational and Fine-Grained Argument Mining.“ In: *Datenbank-Spektrum* (2020), pp. 1–7.

[180] Frans H Van Eemeren, Rob Grootendorst, and Tjark Kruiger. *Handbook of argumentation theory: A critical survey of classical backgrounds and modern studies*. Vol. 7. Walter de Gruyter GmbH & Co KG, 2019.

[181] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. „Attention is all you need.“ In: *Advances in neural information processing systems (NeurIPS)*. 2017, pp. 5998–6008.

[182] Eva Maria Vecchi, Neele Falk, Iman Jundi, and Gabriella Lapesa. „Towards argument mining for social good: A survey.“ In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2021, pp. 1338–1352.

[183] S Villata et al. „Using Argument Mining for Legal Text Summarization.“ In: *Legal Knowledge and Information Systems: JURIX 2020: The Thirty-third Annual Conference, Brno, Czech Republic, December 9-11, 2020*. Vol. 334. IOS Press. 2020, p. 184.

[184] Denny Vrandečić and Markus Krötzsch. „Wikidata: A Free Collaborative Knowledgebase.“ In: *Commun. ACM* 57.10 (Sept. 2014), pp. 78–85. ISSN: 0001-0782. DOI: 10.1145/2629489. URL: http://doi.acm.org/10.1145/2629489.

[185] Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. „Using argument mining to assess the argumentation quality of essays.“ In: *Proceedings of COLING 2016, the 26th international conference on Computational Linguistics: Technical papers*. 2016, pp. 1680–1691.

[186] Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. „Using Argument Mining to Assess the Argumentation Quality of Essays.“ In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 1680–1691. URL: https://aclanthology.org/C16-1158.

[187] Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk 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*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 176–187. URL: https://aclanthology.org/E17-1017.

[188] 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: *Proceedings of the 4th Workshop on Argument Mining*. 2017, pp. 49–59.

[189] Henning Wachsmuth, Martin Potthast, and Matthias Hagen. „Overview of Touché 2021: Argument Retrieval.“ In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings*. Vol. 12880. Springer Nature. 2021, p. 450.

[190] Henning Wachsmuth, Shahbaz Syed, and Benno Stein. „Retrieval of the Best Counterargument without Prior Topic Knowledge.“ In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 241–251. DOI: 10.18653/v1/P18-1023. URL: https://aclanthology.org/P18-1023.

[191] Henning Wachsmuth and Till Werner. „Intrinsic Quality Assessment of Arguments.“ In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 6739–6745. DOI: 10.18653/v1/2020.coling-main.592. URL: https://aclanthology.org/2020.coling-main.592.

[192] Vern Walker, Karina Vazirova, and Cass Sanford. „Annotating Patterns of Reasoning about Medical Theories of Causation in Vaccine Cases: Toward a Type System for Arguments.“ In: *Proceedings of the First Workshop on Argumentation Mining*. Baltimore, Maryland: Association for Computational Linguistics, June 2014, pp. 1–10. DOI: 10.3115/v1/W14-2101. URL: https://aclanthology.org/W14-2101.

[193] Kelly Zhang and Samuel Bowman. „Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis.“ In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. 2018, pp. 359–361.

[194] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. „Aligning books and movies: Towards story-like visual explanations by watching movies and reading books.“ In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 19–27.