Leipzig University
Institute of Computer Science
Degree Programme Computer Science, B.Sc.

Cross Domain Counterargument Retrieval using Large Language Models

Bachelor's Thesis

Janko Götze

1. Referee: Prof. Dr. Martin Potthast

Submission date: December 21, 2023

Declaration

Unless otherwise indicated in the text or references, this thesis is entirely the product of my own scholarly work.

Leipzig, December 21, 2023

. . . .

Janko Götze

Abstract

Computational methods for identifying and generating counterarguments in online discussions are increasingly becoming a focus of research. Even though counterarguments are often associated with persuasion in theoretical and practical argumentation, to the best of our knowledge, this connection has been neglected in computational argumentation. Against this background, the aim of the presented work is to better understand and, if possible, further develop methods for identifying persuasive counterarguments in online discussions. The focus of this bachelor's thesis is the prediction of persuasive counterarguments as well as relevant counterarguments in online discussions. This bachelor's thesis explores the computational identification of persuasive counterarguments in online discussions, with a focus on the Change My View subreddit. The bachelor's thesis is based on the SimDissim, approach from Wachsmuth et al. [2018], a simple counterargument retrieval approach based on similarities. Our research investigates its effectiveness in predicting persuasive counterarguments and its adaptability for improvement. Three experiments were conducted using the Change My View dataset, revealing that the SimDissim performs marginally better than random guessing in predicting persuasive counterarguments. However, it proves effective in argument relevancy prediction, achieving an average normalized discounted cumulative gain (nDCG) of 0.9. The findings contribute to understanding the challenges and potential improvements in computational methods for identifying persuasive counterarguments in online discussions. The results of this thesis showed that the model used is well suited to predict argument relevance. Further experiments would need to investigate whether the model is also suitable for predicting other characteristics of counterarguments.

Contents

Acknowledgements

I thank Dr. Johannes Kiesel and Nailia Mirzakhmedova for their support and guidance during the writing of this thesis. Without their help, this thesis would not have been possible.
I also thank the authors of the webisthesis template for their excellent work!

Chapter 1

Introduction

A counterargument, also known as an opposing argument, is a viewpoint or an argument that opposes or refutes an existing argument or claim. Counterarguments are often used in debates, discussions, persuasive writing, and critical thinking to provide a more comprehensive view on a topic and to address potential objections or weaknesses in an argument.

In natural language processing, counterarguments play a key role in various applications, such as question answering, summarization, and dialogue generation. In these applications, it is often necessary to be able to identify the most salient counterarguments to a given claim, in order to provide a comprehensive and informative response.

Existing research focuses on counterargument detection [Körner et al., 2021], generation [Alshomary et al., 2021] and retrieval [Wachsmuth et al., 2018]. Even though counterarguments are often associated with persuasion in theoretical and practical argumentation, to the best of our knowledge, this connection has been neglected in computational argumentation. One particularly valuable approach has been proposed in the pioneering work on "best" counterargument retrieval by Wachsmuth et al. [2018]. They defined their target for a counterargument as the one that "invokes the same aspects as the (input) argument while having the opposite stance." Their approach to finding the best counterargument is to select an argument that is the most semantically similar to the given argument and simultaneously most dissimilar to it in terms of stance. From here on we will refer to this approach as SIMDISSIM. A more detailed description of SIMDISSIM can be found in Section 3.1. We go into more depth about further existing work in Chapter 2.

Similar to the work discussed above, this thesis aims to identify the best counterarguments in the domain of online discussions, specifically in the Reddit based discussion forum: Change My View. Change My View provides a dynamic and inclusive platform for users to participate in thoughtful discussions by encouraging users to present well-reasoned arguments. The discussions are carefully moderated according to the community rules[1] ensuring the discussions maintain a high standard of quality and remain focused on the topic at hand. According to the rules, direct responses to a submission must challenge or question at least one aspect of the submitted view.

Additionally, Change My View allows users to indicate if they have been persuaded by other user(s)' comment through the delta mechanism, which allows users to highlight the comment(s) that changed their view to other users. These two factors make Change My View suitable both for the analysis of counterarguments and persuasion. This thesis builds upon contributions made by Wachsmuth et al. [2018] and poses the following research questions:

- To what extent can we effectively predict persuasive counterarguments in online discussions using the idea of SimDissim?
- Can SimDissim be adapted or modified to improve its effectiveness in online persuasive discussions?
- Can SimDissim predict other characteristics of counterarguments?

In Chapter 3 we describe the methods chosen to answer these questions and in Chapter 4 we describe the experiments conducted. We carried out three experiments that are based upon SimDissim and the Change My View dataset of Al-Khatib et al. [2020].

The first experiment was focused on the task of persuasive counterargument prediction and employed all comments from the same thread as candidates. In the second experiment, we narrowed down the candidate pool to root comments from the same thread, which are enforced by community rules to challenge the original argument. The third experiment was focused on the task of argument relevancy prediction and employed delta comments from different threads as candidates. The results of the experiments are discussed in Section 4.3.3. The main findings are as follows:

- For persuasive counterargument prediction SimDissim is just $5 - 10\%$ better than random guessing.
- For argument relevancy prediction SimDissim achieves an nDCG of 0.9.
- SimDissim cannot be improved for persuasive counterargument prediction by minor modifications, without changing the underlying idea.

[1] Change My View community rules: `https://www.reddit.com/r/changemyview/wiki/rules`

Chapter 2

Background

To have a foundation for the experiments, we first look at the theoretical background of argumentation including some real-world examples. We will also look at computational argumentation, which includes different approaches to argument retrieval and generation. Finally, we will look at the problem of persuasiveness prediction.

2.1 Argumentation Theory

First let us look at the theoretical background of argumentation. An integral part of argumentation theory is, as the name implies, the definition of an argument. There are multiple different frameworks defining arguments.

One of these frameworks for analyzing and understanding arguments is the Toulmin Model [Toulmin, 2008]. An example of the model is shown in Figure 2.1. It breaks down an argument into essential components. The claim represents the main assertion, the data provides support, and the warrant is the underlying reasoning that connects the data to the claim. Additionally, the model can be extended with backings, qualifiers, and rebuttals to further refine the argument and address counterarguments.

Another framework for analyzing arguments is the Freeman Model [Freeman, 2011]. Both models identify premises, conclusion, and qualifiers in an argument. However, the Freeman Model focuses more one the interactions between a defending and an opposing argument. It addresses conditions of rebuttal, which are the conditions under which the argument would no longer be valid. The Freeman Model is more comprehensive, emphasizing explicit warrants, introducing backing, and addressing conditions of rebuttal, making it a structured approach for in-depth argument analysis, while the Toulmin Model primarily identifies the basic components of an argument.

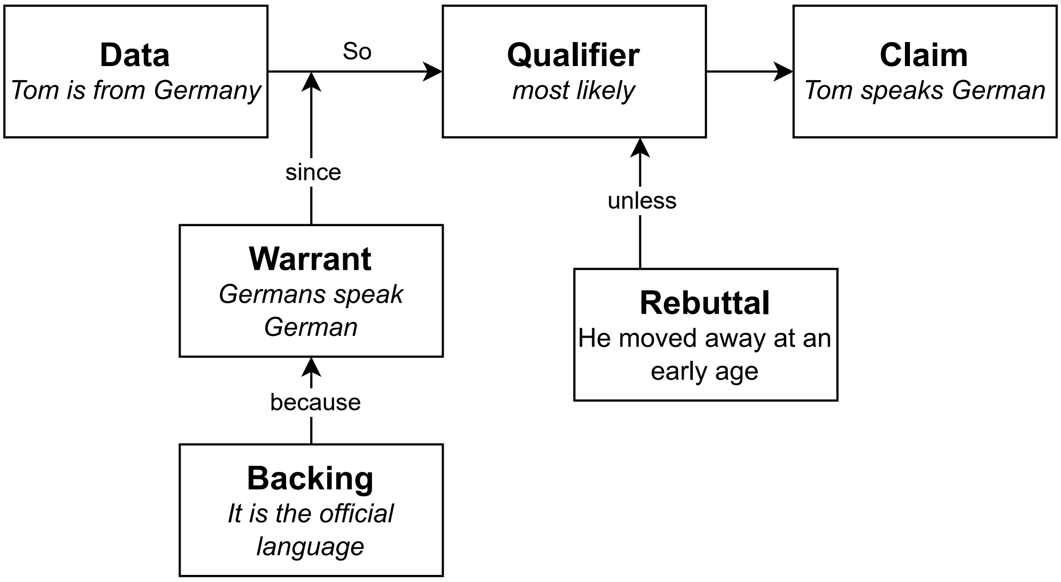


Figure 2.1: Toulmin Model Example: "Tom is from Germany, so he speaks German, since everyone from Germany speaks German, because it is the official language of Germany, unless they moved away at an early age."

One additional concept specific to counterarguments, is *defeasible reasoning* [van Eemeren et al., 1987]. This idea was introduced by Pollock [1987] and is based on the concept of *defeaters*. Defeaters are either rebutting or undercutting. Rebutting defeaters are arguments that directly oppose the original argument. Undercutting defeaters are arguments that weaken the original argument. His standard example is "X is red, because X looks red to me, but X is illuminated by a red light, so X could have another color."

There are also some concepts and structures that only arise when multiple arguments come together and form a discussion. These structures and interactions between individual arguments can lead to new insights and conclusions, that are not present in the individual arguments themselves [Mirzakhmedova et al., 2023]. One example of a concept specific to discussions are fallacies. A fallacy in argumentation is a "deception in disguise" [Habernal et al., 2018]. It is a flaw in reasoning that makes an argument invalid. Fallacies are a common occurrence in arguments and are even more prominent in online discussions [Habernal et al., 2017]. The probably most famous one is the *ad hominem* fallacy. Ad hominem arguments attack the character or personal attributes of their opponent rather than addressing the substance of their argument. It is especially common in digital discourse where participants are anonymous and often resort to personal attacks, name-calling, or character assassinations instead of engaging in rational debate.

2.2 Real-world Argumentation

In this section we will highlight some real-world examples of argumentation. We will particularly focus on online discussions because they are the most accessible for computational analysis.

2.2.1 Change My View

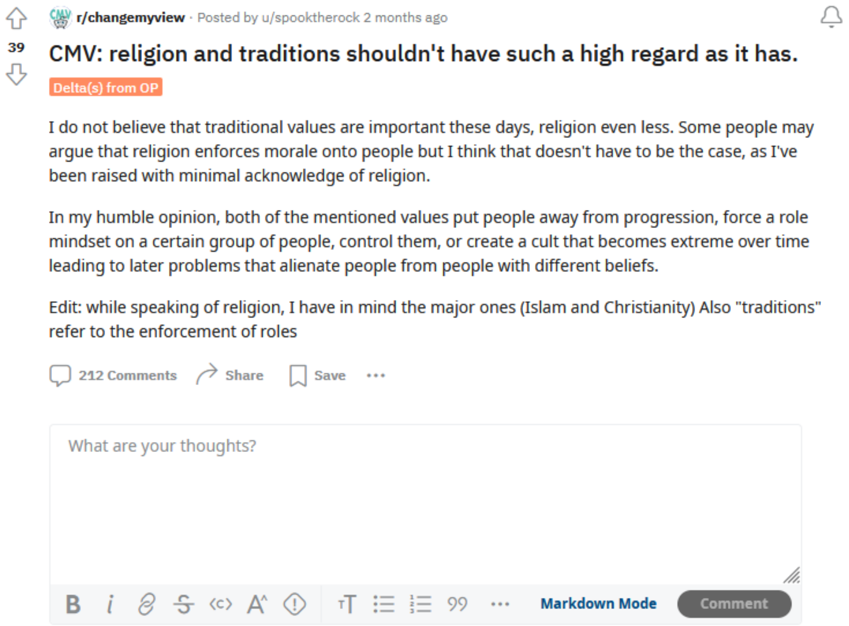
One example of real-world argumentation is *Change My View*, which is part of the online platform Reddit. Reddit is structured in Subreddits, which are like mini forums. Everyone can create threads, and everyone can comment. A thread is a tree with an *original post* (OP) as a root and comments below it. These trees can have a nearly[1] arbitrary depth. An example can be seen in Figure 2.2. Even though the platform is very structured, the users are known for their chaotic and unpredictable behavior. Luckily, this does not hold true for all the Subreddits. Change My View is one of the more structured ones. It has rules and a general structure to the threads, which are enforced by moderators, but still everyone is able to create threads and comment on most topics.

The idea of *Change My View* is to post a controversial opinion and then others can comment on it and try to change the OPs view. When someone is convinced, they can award a delta to the comment, which changed their view. These deltas are visible to everyone so that they can see which arguments are the most convincing. For our purposes, these deltas are considered as a ground truth for the persuasiveness of an argument, as done in Tan et al. [2016] and Al-Khatib et al. [2020]. An example of a Change My View thread with a delta can be seen in Figure 2.3.

2.2.2 iDebate

The iDebate website is another online platform that serves as a practical example of argumentation in the real world. This website offers a space for individuals to engage in structured and informed discussions, often involving controversial or complex topics. Users can present their viewpoints, provide supporting evidence, and engage in debates with others in a respectful and logical manner.

[1] There is a Subreddit called r/counting where users just count, by replying with the next number to the previous comment. At some point an admin had to step in and stop them, because they were causing *site-wide performance issues*. This is the same Subreddit that was responsible for some *glitch tokens* in GPT-3.



r/changemyview · Posted by u/spooktherock 2 months ago

CMV: religion and traditions shouldn't have such a high regard as it has.

I do not believe that traditional values are important these days, religion even less. Some people may argue that religion enforces morale onto people but I think that doesn't have to be the case, as I've been raised with minimal acknowledge of religion.

In my humble opinion, both of the mentioned values put people away from progression, force a role mindset on a certain group of people, control them, or create a cult that becomes extreme over time leading to later problems that alienate people from people with different beliefs.

Edit: while speaking of religion, I have in mind the major ones (Islam and Christianity) Also "traditions" refer to the enforcement of roles

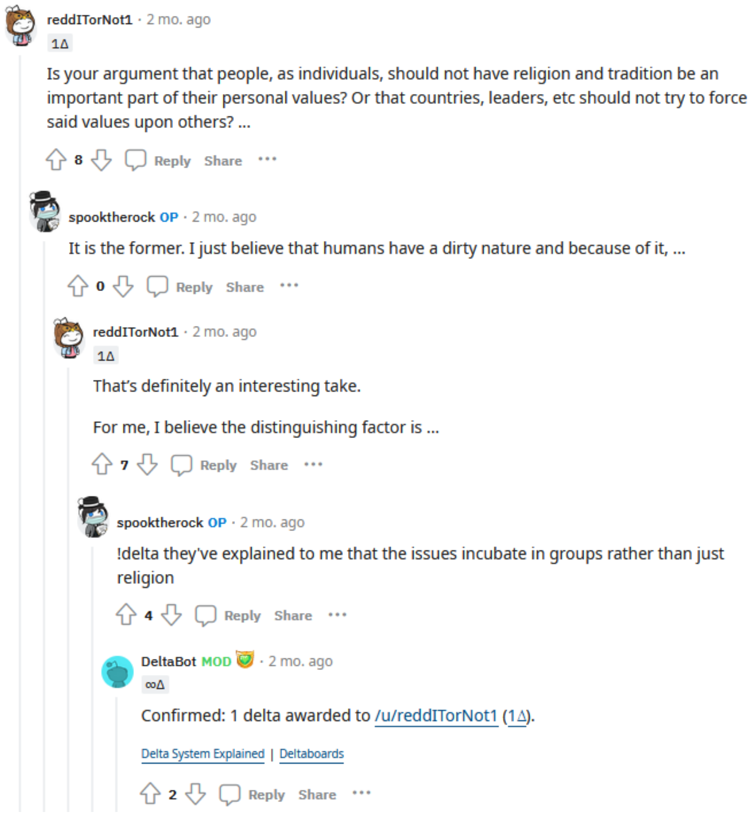
Figure 2.2: Example of a Change My View OP, taken from `https://www.reddit.com/r/changemyview/comments/17ksj3o/cmv_religion_and_traditions_shouldnt_have_such_a/`

The website provides a diverse range of topics and promotes critical thinking and effective communication. It encourages users to develop and present well-structured arguments, cite credible sources, and engage in constructive dialogue. Through this platform, individuals can refine their argumentation skills, gain a deeper understanding of various perspectives, and foster productive discussions on important societal issues.

In contrast to Change My View on Reddit, iDebate is a more structured platform. Topics and arguments are formulated extremely precisely, short and generalized. Also, the stance of the argument is always clear, and every argument has a source. On Change My View people express their opinions in a more personal way, which makes it harder to analyze them computationally, but also more representative for casual everyday discussions.

2.3 Computational Argumentation

Computational argumentation is a field of research that focuses on developing models and methods for analyzing and understanding arguments. It is part of computational linguistics and natural language processing. It covers a lot of tasks from automatic extraction and evaluation of arguments from natural



reddITorNot1 · 2 mo. ago
1Δ

Is your argument that people, as individuals, should not have religion and tradition be an important part of their personal values? Or that countries, leaders, etc should not try to force said values upon others? ...

spooktherock OP · 2 mo. ago

It is the former. I just believe that humans have a dirty nature and because of it, ...

reddITorNot1 · 2 mo. ago
1Δ

That's definitely an interesting take.

For me, I believe the distinguishing factor is ...

spooktherock OP · 2 mo. ago

!delta they've explained to me that the issues incubate in groups rather than just religion

DeltaBot MOD · 2 mo. ago
∞Δ

Confirmed: 1 delta awarded to /u/reddITorNot1 (1Δ).

Delta System Explained | Deltaboards

Figure 2.3: Example of a Change My View thread with a delta, shortened for better readability. Taken from `https://www.reddit.com/r/changemyview/comments/17ksj3o/comment/k79oay/?utm_source=share&utm_medium=web2x&context=3`

language text to argumentation generation [Boltuzic and Snajder, 2014]. In this section, we will look at different parts of computational argumentation, focused on argument and counterargument retrieval.

One significant area within computational argumentation is argument retrieval. Argument retrieval is the process of searching and extracting specific arguments or argumentative content from a dataset or corpus of text [Wachsmuth et al., 2017]. This retrieval aims to find relevant arguments related to a particular topic, debate, or context [Lin et al., 2023]. One common approach to argument retrieval is to first extract a set of candidate arguments and then rank them by their relevance [Green et al., 2021]. Notably, the Touché lab holds yearly shared tasks and competitions for computational argumentation focused around argument retrieval [Bondarenko et al., 2023].

Additionally, to argument retrieval, there is also counterargument retrieval. Counterargument retrieval is the process of finding relevant counterarguments that oppose a given argument or stance [Wachsmuth et al., 2018]. It can require an understanding of the argument, its context and the stance of the argument.

There is also some research for counterargument generation. This process is even more challenging than counterargument retrieval because it requires a deeper understanding of the argument and its context [Ein-Dor et al., 2020]. An example of counterargument generation is Alshomary et al. [2021]. They try to find weak premises in a given argument and then generate counterarguments based on these weak premises, to undermine the original argument.

Most approaches to counterargument retrieval and generation use prior topic knowledge. Wachsmuth et al. [2018] tried to solve the problem of counterargument retrieval without prior topic knowledge, by focusing purely on the semantics of the arguments. They used a combination of semantic and syntactic similarity measures to determine the relevance of counterarguments for a given argument. Their approach is the foundation for our experiments.

2.4 Persuasiveness Prediction

Most of the time, the goal of a discussion is to convince the other participants of your opinion. Especially in online discussions does the most persuasive argument win and not the most relevant one [Dimitrov et al., 2021] [Lukin et al., 2017]. There has been plenty of research on the topic of persuasiveness in online media. A lot of it in the context of propaganda detection [Martino et al., 2020].

Tan et al. [2016] concluded that the persuasiveness of an argument is very dependent on numerous factors. They found that the persuasiveness of an argument is very dependent on the person being persuaded. Also, their data showed, that persuasiveness is less dependent on the argument itself, but more on meta information like arguments that come earlier in the discussion, are lengthier or use italics and bullet points are more likely to be persuasive.

Dimitrov et al. [2021] posed a shared task for detecting persuasive techniques in online pictures and text, mainly for propaganda detection. They propose 22 persuasive techniques. They could be helpful, by detecting them in arguments or using them as a guideline for generating persuasive arguments.

Chapter 3

Methods

Since the field of persuasive counterargument prediction is to our knowledge not well researched, we use the SimDissim approach from Wachsmuth et al. [2018], which was originally used for argument relevancy prediction, and adapt it to our tasks. The two tasks we want to solve are argument relevancy prediction and persuasive counterargument prediction.

For persuasiveness prediction we use counterarguments from the same topic and try to find the most persuasive one. For relevancy prediction we use counterarguments from different topics and try to find the one from the initial topic.

In Section 3.1 we describe the SimDissim approach in more detail. Section 3.2 defines the tasks we want to solve and Section 3.3 describes the evaluation metric we use.

3.1 Approach

3.1.1 SimDissim

All of our experiments are based on the SimDissim approach from Wachsmuth et al. [2018]. Their main idea is to use similarity measures as an indicator for the relevance of counterarguments. More precisely, they hypothesize the best counterargument to invoke the same aspects as the initial argument while having the opposite stance. They model this relation by using a combination of semantic and syntactic similarity measures. The two similarity measures used are the pure words similarity and the embedding similarity. The pure word similarity is the Manhattan similarity of the bag-of-words representation of the arguments. The embedding similarity is the Word Mover's Distance [Kusner et al., 2015] between the word embeddings of the arguments. Additionally, the initial argument is split into premise and conclusion and the similarities are

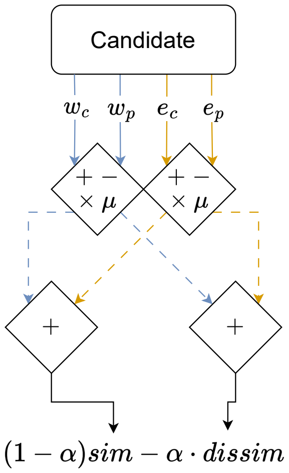


Figure 3.1: General structure of the scoring function: Word similarities in blue, embedding similarities in orange. The rhomboids represent the aggregation functions. Dotted lines represent on possible result of the aggregation functions.

calculated separately for those parts. In the end there are 4 similarities for each counterargument, embedding and pure word similarity for premise and conclusion each. The final score for each counterargument is a combination of the four similarities, we call this the *scoring function*. This score is calculated separately for each counterargument and then the counterarguments are ranked by their score.

3.1.2 Scoring Function

To solve both tasks we use a scoring function inspired from Wachsmuth et al. [2018]. A sketch of the whole scoring function can be found in Figure 3.1. Following SimDissim we calculate similarity scores for words w and embeddings e for both the premise p and the conclusion c for each candidate, w_c, w_p, e_c and e_p respectively. In our experiments we modify the original similarity measures by using the cosine similarity instead of the Manhattan similarity and the Word Mover's Distance. This modification is motivated by the fact that the cosine similarity is more robust to outliers and is less computationally expensive. We then condense these four measures into one word similarity and one embedding similarity, which we refer to as *unit similarities.*

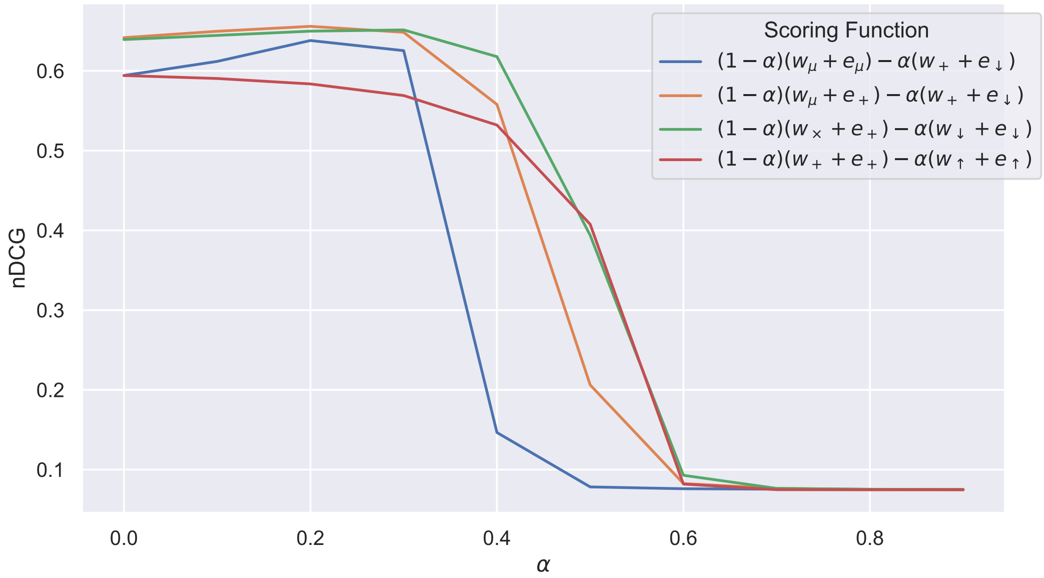


Figure 3.2: Visualization of the nDCGs of the α-tests. Shown are some of the best performing combinations and the one from the original paper(red). Our final choice was green with $\alpha = 0.3$.

For this step, we used the following aggregation functions:

$$\begin{aligned} w_+ &= w_p + w_c \\ w_\times &= w_p \cdot w_c \\ w_\uparrow &= \max(w_p, w_c) \\ w_\downarrow &= \min(w_p, w_c) \\ w_\mu &= \frac{w_p + w_c}{2} \end{aligned} \tag{3.1}$$

The same functions are used for the embeddings with e_+, $e_\times$, $e_\uparrow$, $e_\downarrow$ and e_μ respectively. The general formula for the final score is as follows:

$$score = \alpha \cdot sim - (1 - \alpha) \cdot dissim \tag{3.2}$$

where *sim* and *dissim* are a combination of a word unit similarity and an embedding unit similarity. The α represents how impactful the dissimilarity should be. An α of 1 means that only the similarity is used and an α of 0 means that only the dissimilarity is used.

We tested all combinations of the unit similarities with α values between 0 and 1 in 0.1 increments. A selection of the results can be found in Figure 3.2. The one from the original paper, shown in red, was under the top 20% of all tested metrics. Considering that there were over 600 different combinations of metrics, it was originally used on a completely different dataset, with a different setup, and it still performed better than most of our metrics, shows that SimDissim has some generalization capabilities. The one we will use for the rest of the experiments is *psmin*(product, sum, min):

$$psmin = 0.7(w_{\times} + e_{+}) - 0.3(w_{\downarrow} + e_{\downarrow}) \tag{3.3}$$

3.1.3 Objective of the Original Study

During the initial experimental phase, a disparity emerged between the objective of Wachsmuth et al. [2018] and our own. Even though they called their task “finding the best counterargument to any argument”, we argue that with their setup they did not find the best counterargument, but the most relevant one, and therefore we call it *argument relevancy prediction*. Their study focused on identifying related arguments across different discussion threads, whereas our focus centers on ranking arguments within the same thread. Furthermore, our experimental setup diverged from the original paper in several aspects. Notably, the dataset used in our research differed from theirs, and our model accommodates multiple correct counterarguments as opposed to their approach of a single correct counterargument.

While the notion of similarity serves as a viable indicator for detecting related counterarguments, it is important to acknowledge that the attributes that contribute to a strong counterargument extend beyond mere similarity.

Following this line of thought, we specified two tasks for our experiments. The first one is to find persuasive counterarguments and the second one is to find relevant counterarguments.

3.2 Task Definition

For the relevancy task we have to answer the question: *How persuasive is the counterargument given the initial argument?* The candidates are counterarguments from the same topic as the original argument. We have to rank them by their persuasiveness.

For the relevancy task we have to answer the question: *Does the counterargument fit the original argument or not?* This task is similar to the work of Wachsmuth et al. [2018]. Given one initial argument and multiple counterargument candidates from different discussion threads, the task is to find the candidate which fits the topic of the original argument.

To solve both tasks we use the scoring function from the previous section and rank the candidates by their score.

3.3 Evaluation

For Evaluation, we decided to use the nDCG [Järvelin and Kekäläinen, 2002], which is a ranking metric. It is normalized for the number of correct and total candidates. Since every thread has a different number of true and total candidates, this allows us to directly compare the nDCGs of different threads.

The general idea of nDCG is that each candidate gets less influence on the final score the further down it is ranked. Also, it allows for each individual candidate to have a custom weighting for the influence on the final score. In our case we use 0 for wrong candidate and 1 for correct candidate.

The general formula is as follows:

$$nDCG = \frac{\text{DCG}}{\text{IDCG}} \tag{3.4}$$

Where DCG is the discounted cumulative gain and IDCG is the ideal discounted cumulative gain:

$$\text{DCG} = \sum_{i=1}^{n} \frac{2^{rel_i} - 1}{\log_2(i+1)} \tag{3.5}$$

$$\text{IDCG} = \sum_{i=1}^{N} \frac{1}{\log_2(i+1)} \tag{3.6}$$

The DCG is the sum of the relevance scores of the candidates, where the relevance score of each candidate is discounted by the position in the list. To normalize the DCG we divide it by the IDCG, which is the DCG if all correct candidates are at the top, i.e. the ranking with the highest possible score. It is also common to use the nDCG@k, which is just the nDCG of the top k candidates.

Since the nDCG is just the DCG of our ranking divided by the DCG of the ideal ranking, it can be interpreted as the percentage of the ideal ranking we achieved. If an approach achieved an nDCG of 1, it would mean that it is the best possible approach. If it achieved an nDCG of 0.5, it would mean that it is half as good as the best possible approach.

Chapter 4

Experiments

To evaluate the effectiveness of the SimDissim approach for the two tasks of counterargument persuasiveness prediction and counterargument relevance prediction (see Chapter 3), we applied it to two different real-world argument datasets: the ArguAna corpus [Wachsmuth et al., 2018] and the Webis-CMV-20 dataset [Al-Khatib et al., 2020]. In the following, we describe the datasets and the preprocessing steps we applied to them in more detail. We then describe the experimental setup, the results we obtained and how they answer our research questions.

4.1 Datasets

ArguAna Corpus The ArguAna corpus [Wachsmuth et al., 2018] is an English corpus for studying the retrieval of the best counterargument to a given argument. It contains 6753 pairs of argument and best counterargument from the online debate portal idebate.org. Due to the nature of the debate portal, the corpus contains only one counterargument per argument. For this reason, we only used the ArguAna corpus for the counterargument relevancy prediction task.

Change My View Dataset The Webis-CMV-20 dataset [Al-Khatib et al., 2020] comprises all available posts and comments in the Change My View subreddit from the foundation of the subreddit in 2005, until September 2017. The dataset contains 28,722 unique posts with more than 3 million unique comments.

Task/Setup	Discussions	Comments	Delta Comments	Avg. Comments per Delta
Persuasiveness				
All Comments	19045	1892349	34425	63.95
Root Comments	19096	298143	36489	10.46
Relevancy	11227	11227	11227	Number of Candidates

Table 4.1: Overview of the Data used for the two tasks on the Change My View Dataset. The relevancy prediction task is special, because we condensed every thread into one delta comment and the number of candidates varies depending on the experiment.

4.2 Data Preprocessing

Preprocessing and filtering were essential steps before utilizing the datasets for our analysis. The approach varied for each dataset.

For the ArguAna corpus we only had to convert the data into a format, which our model could process. That included extracting the title and splitting the argument into premise and conclusion.

In contrast, the Webis-CMV-20 dataset required more extensive preprocessing. Originally consisting of raw, unfiltered threads that were collected using the Reddit-API, which included additional information, the dataset was refined to include only relevant details. For the purposes of our work, we only used the title and the content of the original post (OP) and all comments, with their timestamps, up and down votes, and content.

We excluded all deleted and removed threads and comments, as they functioned solely as placeholders without any substantive content. We also excluded all threads where no delta was awarded, since we use the delta-awarded comments as ground truth for the persuasive counterargument prediction task.

To minimize artificial similarities thread titles starting with “CMV:” were trimmed. Furthermore, the model requires the initial argument to be split into premise and conclusion, we use the thread title as the conclusion and the content of the OP as the premise.

The metadata of the comments did not include information about awarded deltas, so we had to extract them manually. A comment is awarded a delta, if someone replies to it with one of a few predefined *delta phrases*. If it is recognized as a valid delta that adheres to the community rules[1] a custom bot for the Subreddit will reply to the comment with a confirmation message. We

[1] https://www.reddit.com/r/changemyview/wiki/rules/

	nDCG					
Approach	all		@10		@5	
Candidate Pool	mean	std	mean	std	mean	std
Our Method						
root comments	0.53	0.25	0.50	0.28	0.41	0.33
all comments	0.39	0.21	0.26	0.29	0.20	0.29
Baseline random						
root comments	0.47	0.25	0.44	0.29	0.34	0.33
all comments	0.29	0.16	0.13	0.22	0.08	0.20
Baseline longest						
root comments	0.64	0.26	0.61	0.30	0.55	0.35
all comments	0.52	0.26	0.42	0.33	0.36	0.35
Baseline earliest						
root comments	0.50	0.21	0.46	0.25	0.36	0.32
all comments	0.42	0.22	0.30	0.29	0.23	0.31

Table 4.2: Persuasiveness Results: Comparison of the nDCG@k scores in relation to the number of candidates, getting harder from top to bottom.

filter for these confirmation messages by the bot to extract the delta comments. In line with prior research, we consider only comments with an awarded and confirmed delta as persuasive. An example of an awarded delta can be seen in Figure 2.3.

4.3 Results

4.3.1 Counterargument Persuasiveness Prediction

Following related work, we used the deltas as a ground truth for the persuasive counterargument prediction task. Over 95% of the arguments are not delta comments, but we assume that most of them are still counterarguments and therefore valid candidates. An overview of the data used for the different tasks can be found in Table 4.1. The task definition is as follows: Given an argument and multiple counterargument candidates taken from the same discussion thread, find the candidate which is the most persuasive. To solve this task, we use our scoring function (see Eq. 3.3) and rank the candidates by their score.

As a second scenario we wanted to simplify the problem and reduce the number of candidates, whilst keeping the number of correct candidates the same thus lowering the count of false candidates in relation to the correct ones.

Instead of using all comments as candidates, we only used root comments, because they are required to challenge the OP by the community rules. If a root comment has any delta comment below it in the discussion tree, it is considered as a delta comment. This way the overall number of candidates is reduced, but the number of delta comments stays the same.

As seen in Table 4.2, using all comments performed better than random guessing, beating it by $5 - 10\%$. Using only root comments yielded comparable results. As further baselines we used approaches that always choose the chronologically first comment and the longest comment, respectively. Over all cases SimDissim was about as good as always taking the chronologically first comment as the best one but was beat by about 10% by always taking the longest comment as the best one. Event though considering only root comments did improve the performance of our model, it also improved the performance of the baseline approaches, keeping the margin about the same. In conclusion, we can say that our model is not able to predict the persuasiveness of counterarguments on the Change My View dataset reliably.

4.3.2 Counterargument Relevancy Prediction

For the relevancy prediction experiments on the Change My View dataset we decided to only consider the delta comments, since they are enforced to be both persuasive and relevant, according to the community rules. As a baseline approach, we only used random guessing, since it is the only one that applies to this task. Furthermore, we selected the best rated delta comment from each thread. The rating is calculated by the number of upvotes minus the number of downvotes. By doing this we can represent each thread just by one comment.

As false candidates random delta comments from other threads are used. It has to be noted that we assume that there are just a handful of threads with the same topic. This is not necessarily true, but we decided that it is not in the scope of the thesis to filter or group the threads by similar topics.

The Results can be found in Table 4.3. We experimented with different numbers of total candidates. As expected, the more candidates there are, the harder it is to find the correct one. But even with all threads as candidates the model still outperforms random guessing by at least 55%. Especially when we look at the nDCG@5, we can see that the model performs up to 90% better than random guessing. Even for a few candidates and no limited nDCG, which is the most generous metric, the model performs about 70% better than random guessing.

	nDCG					
	all		@10		@5	
Candidate Pool	mean	std	mean	std	mean	std
Wachsmuth						
Same Debate						
opposing counters	0.80	0.29	0.83	0.21	0.83	0.21
counters	0.70	0.28	0.71	0.26	0.68	0.32
opposing arguments	0.67	0.29	0.68	0.27	0.64	0.34
all arguments	0.57	0.29	0.55	0.33	0.47	0.40
Same Topic						
opposing arguments	0.63	0.31	0.59	0.37	0.56	0.40
all arguments	0.51	0.31	0.45	0.38	0.40	0.41
Our Method						
100 threads	0.94	0.18	0.94	0.20	0.93	0.21
500 threads	0.88	0.24	0.87	0.27	0.86	0.29
1000 threads	0.84	0.27	0.83	0.31	0.81	0.33
all threads	0.65	0.35	0.61	0.41	0.59	0.43
Baseline random						
100 threads	0.20	0.11	0.04	0.15	0.03	0.14
500 threads	0.14	0.06	0.00	0.07	0.00	0.06
1000 threads	0.12	0.04	0.00	0.05	0.00	0.04
all threads	0.08	0.01	0.00	0.01	0.00	0.01

Table 4.3: Relevancy Results: Comparison of the nDCG@k scores in relation to the number of candidates, getting harder from top to bottom. As expected the more candidates there are, the harder it is to find the correct one. In all tasks our model performs way better than random guessing.

Since the nDCG can be interpreted as the percentage of the best possible score, we can say that the model is about 90% as good as it could be for scenarios with less than 500 candidates. For all candidates it is about 60% as good as it could be. This correlation can be seen in figure 4.1.

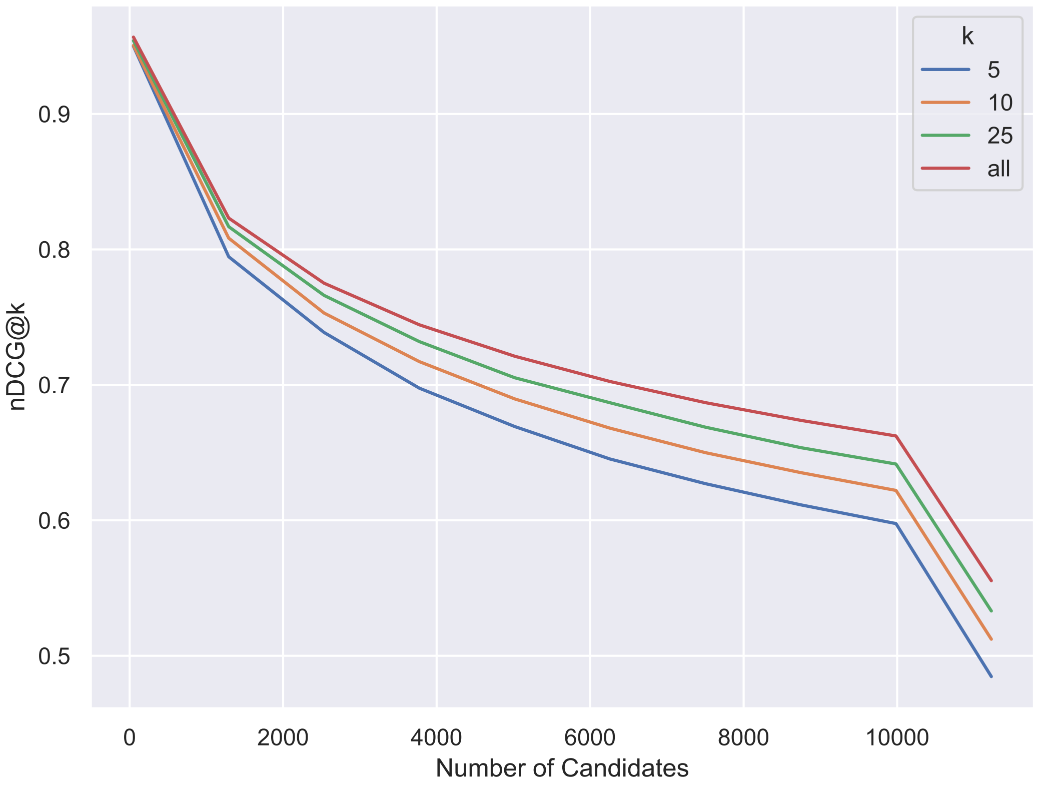


Figure 4.1: Comparison of the nDCG@k scores in relation to the number of candidates with the best performing metric

ArguAna

This scenario is the same as the one from Wachsmuth et al. [2018]. The only difference is that we used our implementation of the model instead of the one from the paper. It has to be noted, that because the algorithm has no training phase, they only used the test set in their experiments, and we did the same. We ran all six different experiments. The candidates for these are:

1. opposing counterarguments form the same debate
2. counterarguments from the same debate
3. opposing arguments from the same debate
4. all arguments from the same debate
5. counterarguments from debates with the same topic
6. all arguments from debates with the same topic

The results can be found in Table 4.3. At the high end, the model gets a nDCG of 0.83 and at the low end it gets a nDCG of 0.40. Again, it can be seen, that the performance of their model gets worse the more candidates there are. Compared to the results from the Change My View dataset, the results are similar, but in general our approach performs about 10% better even for the case with all candidates, which has about double the number of candidates for the Change My View dataset.

4.3.3 Discussion

This section will discuss the results from the previous section and highlight how they answer our research questions. For a more detailed discussion of the limitations of our approach and possible future work, see Chapter 5.

To what extent can we effectively predict persuasive counterarguments in online discussions using the idea of SimDissim? Our results show that SimDissim is not suitable for persuasive counterargument prediction in online discussions. It is only $5 - 10\%$ better than random guessing and gets beaten by simple heuristics like the length of the comment. In retrospect, this is not surprising, because the approach was not designed for this task like we discussed in Section 3.1.3.

Can SimDissim be adapted or modified to improve its effectiveness in online persuasive discussions? We tried to narrow the candidate pool to root comments, to simplify the task. This did improve the results, but only by a few percentages. It also improved the performance of the baselines by a similar amount, which means that the relative performance of SimDissim did not change.

Can SimDissim predict other characteristics of counterarguments? The experiments on argument relevancy prediction show, that SimDissim is suitable for this task. Since the nDCG can be seen as a percentage of the best possible score, the results are way better than we expected. For small candidate pools the results are between 80% and 90% of the best possible score. Even for large candidate pools the results are still around 50% to 60% of the best possible score. The scores of the random baseline below 1% highlight, how hard the task is. We also optimized the scoring function for this task, which improved the results by a few percentages.

Chapter 5

Conclusion

After we have presented our results, we will discuss them in this chapter. We will also discuss open questions, the limitations of SimDissim and possible future work.

5.1 Closing Remarks

We have shown, that SimDissim is not suitable for predicting persuasive counterarguments. Even though this was the original goal of the model, it is not very good at it. At least not in our setup. However, we have shown, that it can predict argument relevancy. After realizing that this task is closer to the original goal of the model, it makes sense, that it performs better at it.

There are a few things, that we could try to improve the model. Probably the biggest thing is the whole concept of the scoring function. We just went with the setup from the original paper, but there are a few things that do not have a clear justification. One of them is splitting the argument in premise and conclusion. In the original paper it is explained with argumentation theory, which states that a counterargument attacks either the premise or the conclusion of an argument [Walton, 2006]. To capture both possibilities, the argument is split into premise and conclusion, both are used to calculate the similarities, and then they are combined. During our experiments we found, that just using the whole argument works and just computing the similarities once works as well. It was just a small test, and it did not quite match the performance of splitting the argument, but it might be worth exploring further. One reason for this finding might be, that we just used the title of the argument as the conclusion. Maybe it would work better, if we used some more sophisticated method to extract the conclusion. Requiring the argument to be split into premise and conclusion also makes it harder to use the model in practice. Even though it is a common practice to split arguments into premise and conclusion, if the model also works without it, it would be easier to use.

Another one is that *sim* and *dissim* from the scoring function (Eq. 3.2) are always a sum of a word unit similarity and an embedding unit similarity. Why is that the case? Why not use only embedding unit similarities? The decisions made in the original paper make sense, do not seem unreasonable and produce satisfactory results, but other alternatives do not seem to be explored. Maybe there is a better way to calculate the similarities. Maybe there is a better way to combine the similarities. There are a lot of possibilities, which could be explored. We briefly explored the idea of using a model to classify the stance of the counterargument and then use that as our dissimilarity. But it diverged too much from the original idea of SimDissim, so we did not pursue it further. To keep the scope of this thesis manageable, we did try to keep the model as close to the original as possible and limit exploration of the more radical changes to the underlying idea.

Another idea that came up during the experiments was to use a completely different approach to persuasiveness prediction. One could be to use specific models to compute sentence similarities instead of just counting words. Another one could be to use multiple models. For example, one model to dissect the argument into individual units that are then used by another model determine the persuasiveness. There are also some patterns, that we repeatedly found in delta comments. Citing the original argument, breaking it down into smaller parts, using bulltepoints, etc. Maybe scoring an argument on its general quality based on these patterns and then combining it with counterargument specific scores could work, because for a counterargument to be persuasive, it helps to be of high quality as well. There are a lot of possibilities, which could be explored.

5.2 Limitations

During the experiments, we only tested our model on two datasets. We can only assume how well it performs on other datasets. Since the setup from the original paper performed well on the Change My View dataset out of the box, it indicates that the model has some generalization capabilities, but more testing is needed to be sure. Also, we prefiltered the dataset, and condensed threads to single arguments. This might have introduced some bias.

Additionally, we made some assumptions, which might not be true in general. We assumed that there are no duplicate topics. This is not necessarily true and not realistic. But we do not know how much it affects the results. Lastly, we assumed that all candidates are counterarguments. This is not true, because we did not filter out supporting arguments. In other datasets this might be a bigger problem, but in the Change My View dataset, the community rules helped to systematically filter out most supporting arguments.

5.3 Future Work

The challenges for future research in this area arise from the limitations mentioned. Further investigations would have to clarify to what extent, the model used can be transferred to other datasets. The splitting of the argument into premise and conclusion requires further critical examination.

The results of this bachelor's thesis showed that the model used is well suited to predict argument relevance. Further experiments are needed to investigate whether the model is also suitable for predicting other characteristics of counterarguments.

Bibliography

Khalid Al-Khatib, Michael Völske, Shahbaz Syed, Nikolay Kolyada, and Benno Stein. Exploiting Personal Characteristics of Debaters for Predicting Persuasiveness. In *58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 7067–7072. Association for Computational Linguistics, July 2020. URL `https://www.aclweb.org/anthology/2020.acl-main.632`.

Milad Alshomary, Shahbaz Syed, Arkajit Dhar, Martin Potthast, and Henning Wachsmuth. Argument Undermining: Counter-Argument Generation by Attacking Weak Premises. In *The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, pages 1816–1827. ACL-IJCNLP, August 2021. doi: 10.18653/v1/2021.findings-acl.159. URL `https://aclanthology.org/2021.findings-acl.159`.

Filip Boltuzic and Jan Snajder. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argument Mining, hosted by the 52nd Annual Meeting of the Association for Computational Linguistics, ArgMining@ACL 2014, June 26, 2014, Baltimore, Maryland, USA*, pages 49–58. The Association for Computer Linguistics, 2014. doi: 10.3115/v1/w14-2107. URL `https://doi.org/10.3115/v1/w14-2107`.

Alexander Bondarenko, Maik Fröbe, Johannes Kiesel, Ferdinand Schlatt, Valentin Barriere, Brian Ravenet, Léo Hemamou, Simon Luck, Jan Heinrich Reimer, Benno Stein, Martin Potthast, and Matthias Hagen. Overview of Touché 2023: Argument and Causal Retrieval. In Avi Arampatzis, Evangelos Kanoulas, Theodora Tsikrika, Stefanos Vrochidis, Anastasia Giachanou, Dan Li, Mohammad Aliannejadi, Michalis Vlachos, Guglielmo Faggioli, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 14th International Conference of the CLEF Association*

(CLEF 2023), Lecture Notes in Computer Science, Berlin Heidelberg New York, September 2023. Springer.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. Semeval-2021 task 6: Detection of persuasion techniques in texts and images. *CoRR*, abs/2105.09284, 2021. URL `https://arxiv.org/abs/2105.09284`.

Liat Ein-Dor, Eyal Shnarch, Lena Dankin, Alon Halfon, Benjamin Sznajder, Ariel Gera, Carlos Alzate, Martin Gleize, Leshem Choshen, Yufang Hou, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. Corpus wide argument mining - A working solution. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7683–7691. AAAI Press, 2020. doi: 10.1609/AAAI.V34I05.6270. URL `https://doi.org/10.1609/aaai.v34i05.6270`.

James B. Freeman. Argument structure: Representation and theory. In *Argumentation Library*, 2011. URL `https://api.semanticscholar.org/CorpusID:13831830`.

Tommaso Green, Luca Moroldo, and Alberto Valente. Exploring bert synonyms and quality prediction for argument retrieval. In *Conference and Labs of the Evaluation Forum*, 2021. URL `https://api.semanticscholar.org/CorpusID:237298056`.

Ivan Habernal, Raffael Hannemann, Christiana Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. Argotario: Computational argumentation meets serious games. *ArXiv*, abs/1707.06002, 2017. URL `https://api.semanticscholar.org/CorpusID:29489521`.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 386–396. Association for Computational Linguistics, 2018. doi: 10.18653/v1/n18-1036. URL `https://doi.org/10.18653/v1/n18-1036`.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4): 422–446, 2002.

Erik Körner, Gregor Wiedemann, Ahmad Dawar Hakimi, Gerhard Heyer, and Martin Potthast. On Classifying whether Two Texts are on the Same Side of an Argument. In *26th Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*, pages 10130–10138. Association for Computational Linguistics, November 2021. doi: 10.18653/v1/2021.emnlp-main.795. URL `https://aclanthology.org/2021.emnlp-main.795/`.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. From word embeddings to document distances. In *International Conference on Machine Learning*, 2015. URL `https://api.semanticscholar.org/CorpusID:14674248`.

Jiayu Lin, Rong Ye, Meng Han, Qi Zhang, Ruofei Lai, Xinyu Zhang, Zhao Cao, Xuanjing Huang, and Zhongyu Wei. Argue with me tersely: Towards sentence-level counter-argument generation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16705–16720, Singapore, December 2023. Association for Computational Linguistics. URL `https://aclanthology.org/2023.emnlp-main.1039`.

Stephanie M. Lukin, Pranav Anand, Marilyn A. Walker, and Steve Whittaker. Argument strength is in the eye of the beholder: Audience effects in persuasion. *ArXiv*, abs/1708.09085, 2017. URL `https://api.semanticscholar.org/CorpusID:2586121`.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. A survey on computational propaganda detection. *ArXiv*, abs/2007.08024, 2020. URL `https://api.semanticscholar.org/CorpusID:220483038`.

Nailia Mirzakhmedova, Johannes Kiesel, Khalid Al-Khatib, and Benno Stein. Unveiling the power of argument arrangement in online persuasive discussions. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15659–15671, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.1048. URL `https://aclanthology.org/2023.findings-emnlp.1048`.

John L. Pollock. Defeasible reasoning. *Cogn. Sci.*, 11(4):481–518, 1987. doi: 10.1207/S15516709COG1104_4. URL `https://doi.org/10.1207/s15516709cog1104_4`.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 613–624, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee. ISBN 9781450341431. doi: 10.1145/2872427.2883081. URL `https://doi.org/10.1145/2872427.2883081`.

Stephen E. Toulmin. *The Uses of Argument, Updated Edition.* Cambridge University Press, 2008. ISBN 9780521534833.

Frans H. van Eemeren, Bart Garssen, Erik C. W. Krabbe, A. Francisca Snoeck Henkemans, Bart Verheij, and Jean H. M. Wagemans. Handbook of argumentation theory. In *Handbook of Argumentation Theory*, 1987. URL `https://api.semanticscholar.org/CorpusID:13820363`.

Henning Wachsmuth, Benno Stein, and Yamen Ajjour. "pagerank" for argument relevance. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, pages 1117–1127. Association for Computational Linguistics, 2017. doi: 10.18653/v1/e17-1105. URL `https://doi.org/10.18653/v1/e17-1105`.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. Retrieval of the best counterargument without prior topic knowledge. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 241–251. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1023. URL `https://aclanthology.org/P18-1023/`.

Douglas N. Walton. *Fundamentals of critical argumentation.* Critical reasoning and argumentation. Cambridge University Press, 2006. ISBN 978-0-521-53020-0.