Representing and Determining Argumentative Relevance in Online Discussions: A General Approach

Zhen Guo[1*] and **Munindar P. Singh**[2]

[1] eBay Inc.
[2] North Carolina State University
zguo8@alumni.ncsu.edu, singh@ncsu.edu

Abstract

Understanding an online argumentative discussion is essential for understanding users' opinions on a topic and their underlying reasoning. A key challenge in determining the completeness and persuasiveness of argumentative discussions is to assess how well arguments under a topic are connected. Online argumentative discussions, in contrast to essays or face-to-face communication, challenge techniques for judging argument relevance because online discussions involve multiple participants and often exhibit incoherence in reasoning and inconsistencies in writing style.

We define relevance as the logical and topical connections between small texts representing argument fragments in online discussions. We provide a corpus comprising pairs of sentences, each labeled with argumentative relevance between the sentences in it. We propose a computational approach relying on content reduction and a Siamese neural network architecture for modeling argumentative connections and determining argumentative relevance between texts.

Experimental results indicate that our approach is effective in measuring relevance between arguments, and outperforms strong and widely adopted baselines. Further analysis demonstrates the benefit of using our argumentative relevance encoding on a downstream task—predicting how impactful an online comment is to a certain topic—over an encoding that does not consider a logical connection.

Introduction

Online argumentative discussions are often unproductive due to repetitive or discursive information (Risch and Krestel 2020). They raise challenges in understanding people's opinions and how they relate to each other. Whereas most existing work investigates a specific aspect of online arguments, such as classifying argument components or predicting convincingness (Guo, Zhang, and Singh 2020), our work aims at a fundamental and general problem in computational argumentation and information retrieval: relevance modeling (Fan et al. 2021). Accordingly, we propose a general measurement and representation of relevance between arguments from both topical and logical perspectives.

We posit the notion of ***argumentative relevance*** to represent how a sentence relates to a preceding sentence in an argumentative discussion. Given a topic, suppose a first sentence and a second sentence arise under that topic in an online discussion. Then, the second sentence is *argumentatively relevant* to the first sentence if the second sentence (1) has a reasoning connection to the first sentence and (2) addresses a specific point raised by the first sentence.

Argumentative relevance is beyond just topical relevance, which involves specificity (Dessalles 2016; Durmus, Ladhak, and Cardie 2019a), cogency (Wachsmuth et al. 2017a; Lauscher et al. 2020), and inheritance (Dung and Son 2001). Therefore, measuring argumentative relevance raises new challenges beyond existing semantic similarity measurement, such as general sentence embeddings, which cannot distinguish whether two sentences are purely topically relevant or are relevant in both topical and logical senses.

Modeling and determining the relevance of arguments in online discussions brings up challenges beyond those for arguments in other forms. In contrast with essays and scientific reports in which arguments are constructed by a single author and are elaborate and coherent, online argumentative discussions are open dialogues where texts are casually produced by multiple interlocutors. Arguments in online discussions are short and comprise informal texts with few transition clues. Therefore, determining and modeling argumentative relevance between social media texts goes beyond argument mining from essays. Section describes key traits of online arguments further with examples.

Measuring relevance between arguments can facilitate downstream tasks such as assessing the quality of arguments. Whether a topic or issue is argued in depth can be determined by whether an argumentation scheme (Walton and Reed 2003) is sufficiently fulfilled in a discussion. A chain of reasoning is an important dimension for evaluating arguments. In many online platforms, although every argument is related to a topic, the chain of reasoning (i.e., how these arguments are related to each other and form a chain of reasoning) remains understudied (Schneider, Davis, and Wyner 2012). On social media, if a post is argumentatively relevant to (e.g., supported or opposed by) multiple other posts, it may be considered a sufficiently discussed point. In this case, identifying whether two pieces of text show argumentative relevance is a fundamental step toward identifying chains of reasoning and thus argument evaluation.

We address the following research questions in this work.

[*] Work performed while at North Carolina State University.

Measuring How can we identify relevant arguments—i.e., those that are coherent in reasoning and relevant in topic—from argumentative discussions on social media?

Representing How can we computationally detect and represent relevance between arguments?

Applying How does measuring argumentative relevance facilitate other tasks where the topical and logical relevance of arguments matters?

Motivating Examples and Traits

We draw posts from an online debate platform, Kialo,[1] as our examples. In each example below, *Sent.A* and *Sent.B* are from two posts by different interlocutors.

Trait 1: Distribution Arguments involve reasoning across multiple textual segments.

Unlike in the essay setting, in online argumentative discussions, segments representing different components of an argumentation schema may appear distributed across different posts, possibly originating from different interlocutors. Although each segment individually is not an argument, the segments together form argumentative structures under their specific context.

- *Sent.A* [Men and women are different.]
- *Sent.B* [Men and women have different brains and are genetically different.]

Neither *Sent.A* nor *Sent.B* is an argument but rather a claim if examined individually. However, *Sent.B* is a premise of *Sent.A* under the topic "Gender stereotyping children needs to stop." In such cases, argument mining needs to go beyond the sentence level to retrieve and relate the two sentences.

Trait 2: Implicitness Linguistic indicators for logical connectives are implicit in turn-taking online arguments.

An argument is conventionally structured as having a main claim (i.e., conclusion) and a set of premises leading to the conclusion. Depending on the specific argumentation schema, an argument may potentially include components such as a warrant or a rebuttal (Toulmin 1958). However, identifying arguments and their relations across interlocutors—especially in online forums such as Reddit *Change My View*,[2] Twitter,[3] or Kialo—raise new challenges beyond argument mining from essays. One challenge is that linguistic indicators for logical connectives such as "for example" or "therefore" are often missing in online arguments, and turn-taking is informal. For example, *Sent.B* is a rebuttal of *Sent.A* even though it lacks explicit cues.

- *Sent.A* [There is no way an omnipotent all-loving god could co-exist with evil.]
- *Sent.B* [Karma explains why is there evil alongside with loving God.]

Trait 3: Coherence Topical relevance between two posts does not necessarily indicate argumentative relevance.

Another challenge is that argumentative relevance considers both reasoning and topical connections, whereas existing measurements emphasize topic similarities and, in many cases, overlook reasoning connections between sentences. Below are two sentences extracted from a discussion about "Democrats should not cooperate with Donald Trump's presidency."

- *Sent.A* [Refusing to cooperate with President Trump violates democratic norms and principles.]
- *Sent.B* [Disapproval of Trump is not the same as not wanting cooperation.]

The two sentences contain words "Trump" and "cooperation" and similar words such as "refuse" and "disapprove." Therefore, they appear topically highly similar. However, they are not argumentatively relevant.

Based on the above observations, we propose a new challenge, namely, measuring the relevance of online arguments.

Related Work

We now discuss techniques used for recognizing and representing arguments and for text representation in general. We focus on applications and limitations of those techniques pertaining to the demonstrated motivations.

Linguistic indicators and neural-network-based encoding methods are two paradigms of representation techniques widely used in argument mining. For instance, Tan et al. (2016) and Persing and Ng (2017) measure relations between pairs of comments using similarity estimated based on word-overlap. Habernal and Gurevych (2017) propose TF-IDF of unigrams and bigrams, POS, and other representations for argumentation mining.

Text embeddings are widely used in argument mining and other information-retrieval tasks. In particular, transformer-based models yield top performance in natural language processing tasks, including on argumentation tasks. For instance, Zhao et al. (2021) leverage BERT and topic embeddings to predict persuasiveness of arguments.

Despite the development of sentence embeddings pretrained for a general purpose and fine-tuned for argument tasks, few existing methods address measuring relevance between arguments from multiple interlocutors from both topical and logical perspective. Jo et al. (2018) examine a neural architecture with an attention measurement for attackable sentences. However, they focus on the sentence level and do not represent relations at the sub-sentence (i.e., text segments) level. In contrast, we posit that how text segments are logically connected between sentences is important for measuring argumentative relevance, and we examine such connections in this work.

Some works incorporate additional considerations such as contexts, dynamic features, and additional hand-curated features for argument mining and other tasks such as persuasion. For instance, research has considered argument (positional) structure (Li, Durmus, and Cardie 2020; Stab and Gurevych 2017). Some studies concentrate on identifying argumentation-schema-based components such as claims and premises (Hidey et al. 2017), and motions and assertions (Persing and Ng 2017). Computational methods used in these studies are covered by the aforementioned linguistic features or embedding methods (Cabrio and Villata 2018). Therefore, we consider those studies as potential application

[1] https://www.kialo.com/

[2] https://www.reddit.com/r/changemyview/

[3] https://twitter.com

Pair type	Connection	Text
P-C	Definition clarification	Children have a right to learn in a religiously neutral setting. There is no such thing as a religiously neutral setting.
P-C	Analogy	Gender is an important tool for social organisation. Tools of social organization are really tools of social control.
P-C	Reasoning (undercut)	Every level of football is on some level exploitative. If adult players consent to play on a team, then they are not being exploited.
A-D	– (topical only)	Health care should remain a privilege, not a right. This definition could directly replace "foo" with "health care" and work well.
R-R	– (topical only)	Vegan diets also contain a lot less heavy metals and pollutants. The meat production industry is itself unethical.

Table 1: Examples of relation and overlap between an argument (Arg) and its counterargument (CntArg). We use straight underlines to mark topical relevance and dashed underlines to mark logical relevance. Here, P-C is a parent-child pair; A-D is an ancestor-descendant pairs; and R-R is a pair of randomly selected sentences.

domains of our work. In Section , we compare our approach with Durmus, Ladhak, and Cardie's (2019b), which uses linguistic features and general embeddings.

We discuss additional related work in Section .

ARM: Argumentative Relevance Measurement

Motivated by the above observations and needs, we target this specific challenge: *Given a topic, model the argumentative relations between two sentences and decide whether one sentence is argumentatively relevant to the other.*

To differentiate relevance from textual similarity, we define relevance as requiring that one sentence complement the other sentence in argumentative structures. Such structures include any relations and logical connectives between argument components—whether *Sent.A* is a premise, assertion, evidence, or qualifier of *Sent.B*. Two sentences that are argumentatively relevant to each other should have implicit or explicit reasoning connections.

We focus on modeling argumentative relevance and deciding whether one sentence is relevant to another. We propose a model, dubbed ARM, and train a computational representation for argumentative relations of pairs of sentences.

Data: Paired Arguments

We draw a dataset for paired arguments from Kialo. On Kialo, one discussion represents a topic of interest, and arguments about the same topic (i.e., within the same discussion) are constructed in a tree structure.

One helpful characteristic of Kialo is that each user is asked to reply with an argument that directly supports or opposes its parent claim. If a claim a user posts fits better in another path of reasoning, the claim may be moved by another user to the better-suited place. Kialo has recently garnered attention in computational research on argumentation because of its high-quality arguments and elaborate argument trees (Durmus, Ladhak, and Cardie 2019a,b; Jo et al. 2020; Skitalinskaya, Klaff, and Wachsmuth 2021).

Our objective is to decide if one sentence is argumentatively relevant to another given that both sentences are drawn from the same discussion. The stance (i.e., supporting or attacking) of a post to its parent is marked as *pro* or *con*. Although stance is binary, the tree structure of a discussion is not binary since there may be two or more parallel arguments with the same stance. We first pair and categorize two posts in a discussion by their positional relations. An illustration of pair types is shown in Figure 1.

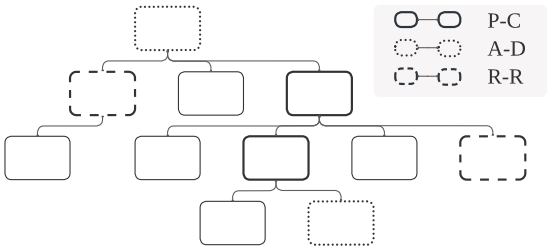

Figure 1: Post-pair types by positional relations.

P-C. Parent-child sentence pairs are extracted from adjacent posts. In a Kialo discussion tree, if two posts are adjacent (with 1-hop distance), we take them as a P-C pair where the first post is the parent and the second post is the child.

A-D. Ancestor-descendant sentence pairs are extracted from posts that belong to the same branch of a discussion tree are but not adjacent. In an A-D pair, the first post is the ancestor post.

R-R. Random pairs. These sentences are from the same discussion but do not belong to a P-C or A-D pair, and are not siblings (i.e., not children to the same parent post), we extract them as R-R pairs.

We extract 1,041 discussions (with 127,040 posts in all) covering diverse topics such as politics, gender, society, technology, and religion. From the discussions, we extract 60,747,130 post pairs, among which 126,103 are P-C pairs. We generate sentence-level pairs from post-level pairs based

on the positions of posts in a discussion tree. Intuitively, P-C pairs are high-relevance pairs whereas A-D and R-R are low-relevance pairs. However, to ensure that relevance labels are reasonable, we incorporate the following additional considerations:

1. A root post on Kialo is always a one-sentence statement and is so general that every other post in the same discussion is argumentatively related to it. Thus, there is no apparent difference in relevance between P-C, A-D, and R-R pairs involving a root post.
2. In A-D pairs, if Post 1 and Post 2 have a two-hop distance (i.e., a grandparent and grandchild pair), the two posts may be argumentatively relevant especially when the grandparent-(parent)-grandchild relation fits in an argument-(counterargument)-refutation structure.
3. Because the specificity of claims along the same path increases with depth (Durmus, Ladhak, and Cardie 2019a) and the focus point shifts in claims with high specificity, claims from different paths may not even be talking about the same topic.

Therefore, we exclude pairs involving a root post and A-D pairs whose distance equals two. To generate pairs at the sentence level and on the same topic, we compare each sentence in Post 1 with each sentence in Post 2 and extract sentences from all possible combinations of sentence pairs if they have at least one common word after stemming and removing stop words. In this way, we end up with 126,103 P-C, 552,660 A-D, and 59,389,604 R-R sentence pairs. Below, for brevity, P-C, A-D, and R-R refer to sentence, and not post, pairs. Table 1 provides examples showing topical and logical connections in different types of pairs.

We conducted a small-scale human study with colleagues to validate the systematically generated labels for relevance. We randomly selected 50 relevant pairs (i.e., P-C pairs) and 50 putatively irrelevant pairs (30 R-R pairs and 20 A-D pairs). These 100 pairs contain 22 pairs that are grandparent-grandchild pairs, root-level arguments, or same-level pairs. The 100 pairs were presented to three raters: an author of this paper, a graduate student with knowledge of argumentation theory, and a graduate student unfamiliar with argumentation theory. The average accuracy score of the three raters was 0.79 against the generated labels. The inter-rater reliability using Krippendorff's alpha for all raters was 0.53. However, when looking at pairwise raters' agreement, Cohen's Kappa coefficient was 0.70 between the author and the rater who is familiar with argumentation, while Cohen's Kappa was 0.40 for the rater unfamiliar with argumentation and the other two raters.

The above study substantiates our above claim as to why those samples are confusing for deciding relevance. Before removing those pairs, the average accuracy of the three raters was 0.73, whereas the average accuracy was 0.79 after removing those 22 pairs. For 88% of the samples, at least one rater agreed with the auto-generated labels.

Although the order of the two sentences in P-C pairs and A-D pairs matters when extracting them from a tree-structured discussion, we consider argumentative relevance as an undirected relation. Properties such as symmetry and transitivity are emphasized in formal approaches, where the argument is constructed in a formal context by a single party. However, these properties are not quite as apparent or natural in real-life multiparty argumentation. For example, in a discussion about "whether science leaves room for free will," *Sent.A* [scientific skepticism allows the possibility of free will] and *Sent.B* [science does require something like the principle of reason] are counterarguments of each other and can be parent-child or child-parent pairs.

Furthermore, tree-structured positional information is specific to Kialo but may not apply on other online forums. As the aforementioned *Trait 1: Distribution* examples show, although *Sent.A* and *Sent.B* form a parent-child pair on Kialo, *Sent.A* may appear after *Sent.B* on some other platform where *Sent.A* is considered a paraphrase of *Sent.B* i.e., indicating agreement with *Sent.B*. Such formal properties are recognized as challenging research problems for tasks such as natural language inference (Schluter and Varab 2018; Wang, Sun, and Xing 2019). We leave them as out of the scope of this paper.

Argumentative Relevance Encoding

We aim to encode a pair of sentences (*sent$_A$* and *sent$_B$*) with an emphasis on their argumentative relevance, i.e., topical and reasoning connections.[4] We capture encoding via binary classification: positive label $y = 1$ means *sent$_A$* and *sent$_B$* are argumentatively relevant; $y = 0$ means they are not. In addition, we examine the performance of our model in generating embeddings for use in downstream tasks such as determining the impact or persuasiveness of an argument.

Our encoding method consists of a content reduction module and a Siamese network architecture. In what follows, we employ simple classifiers and demonstrate the ability of the proposed encoding in downstream tasks.

Content reduction. To capture argumentative relevance, we would like the encoding model to emphasize both topical and reasoning connections, and deemphasize detailed complementary content. Therefore, we propose a content-reduction process to identify such connections and extract connection-related content. To this end, we incorporate dependency parsing and semantic relation search. This process takes word tokens of original sentences as input: $X_A = (x_{A1}, \ldots, x_{An})$ and $X_B = (x_{B1}, \ldots, x_{Bm})$. We use a complementary position mask for each sentence initialized with zeros, i.e., $p_A = (0_{A1}, \ldots, 0_{An})$ and $p_B = (0_{B1}, \ldots, 0_{Bm})$, to mark tokens to be extracted as 1 and tokens to be ignored as 0.

For semantic relations between X_A and X_B, we consider similarity in word usage, analogical syntactical structure, synonyms, and antonyms. To implement this intuition, we use exactly matched stemmed words and adopt WordNet for synonyms and antonyms. If a token x_{Ai} or x_{Bj} from a sentence has such a related word (e.g., the same word or a synonym) in the other sentence, we respectively mark position p_{Ai} or p_{Bj} as 1. We term such tokens *connection tokens*.

In addition, we mark any one-hop tokens from the connection tokens in the corresponding dependency tree as 1.

[4] We use lowercase "sent$_A$" and so on as variables.

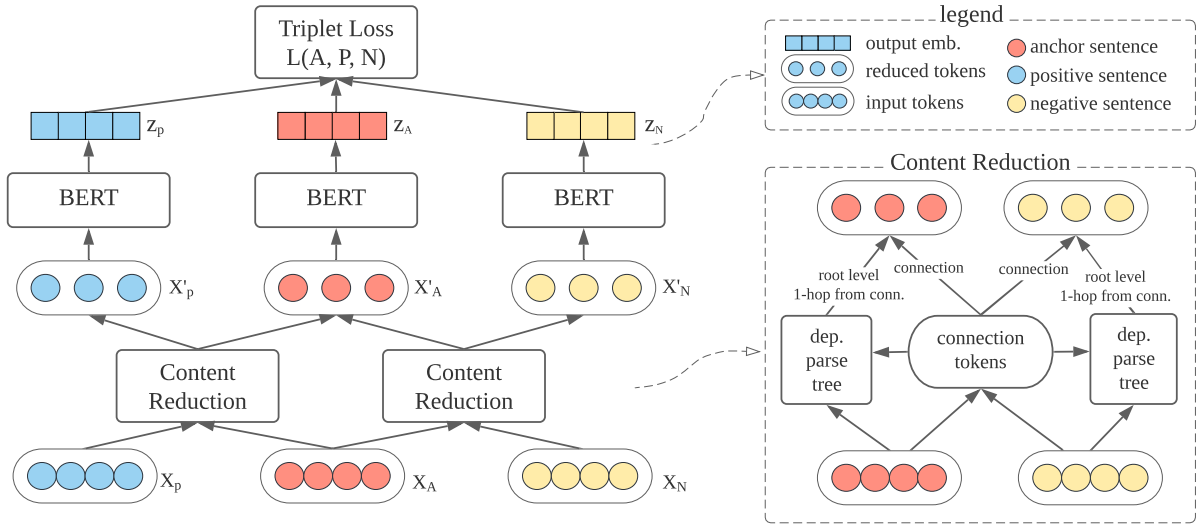

Figure 2: Model architecture: Siamese network with content reduction. The content reduction module applies to both positive (red and blue) and negative pairs (red and yellow). Here, *dep.* means dependency.

The intuition is to capture and retain content that serves a syntactic function related to argumentative structures. For example, "if" and "then" would not be captured by WordNet since WordNet considers nouns, verbs, adjectives, and adverbs; however, an "if…then" clause indicates a proposition and should be captured if the condition or the result being referred is related to a topic of interest.

To preserve the basic structure of a sentence, we extract first-level and second-level tokens from a dependency parse tree. For instance, if a token is a root, one-hop away from the root, or two-hop away from the root, we mark its position in p_A or p_B as 1.

Lastly, we extract tokens marked 1 from the original sentences in their original order to form the reduced content.

This process yields a sequence of tokens of the reduced content for both sentences $X'_A = (x'_{A1}, \ldots, x'_{An})$ and $X'_B = (x'_{B1}, \ldots, x'_{Bm})$.

Below, we illustrate texts before and after content reduction. Tokens in square brackets are irrelevant content being eliminated. Bold tokens are connection tokens. For the remaining preserved tokens, Figure 3 shows their selection based on dependency parse.

- Sent.A ***Human*** *life has a higher value because* ***humans*** *need and depend [on each other], [socially] and [economically].*
- Sent.B ***Humans*** *are the counterpart of [god], provided with [an immortal soul].*

Siamese network. To distinguish relevance from topic similarity, we form positive and negative samples and use triplet loss to train a Siamese network.

Given a sentence A as a sequence of tokens X_A, a positive pair comprises an anchor sentence X_A and another sentence P (tokens X_P), where P is relevant to A. For the same anchor sentence, a negative pair is A and N, where N is a sentence with the same topic that is not argumentatively relevant to A. We use *P-C* pairs from Kialo as positive pairs and *A-D* and *R-R* pairs as negative pairs. We compare parent sentences in *P-C* pairs and the sentence with an earlier timestamp in the *A-D* and *R-R* pairs to locate an anchor sentence and form a triplet training sample (A, P, N).

After content reduction, we feed X'_A, X'_P, and X'_N into a Siamese network. Siamese networks have been demonstrated as effective for STS (Semantic Textual Similarity) and NLI (Natural Language Inference) tasks with BERT as the shared encoding model (Reimers and Gurevych 2019a). We adopt the BERT encoding architecture as the shared model for the Siamese architecture. We use the mean of output embeddings for pooling since the pooling method results in the best performance for textual similarity tasks (Reimers and Gurevych 2019a).

Based on our observations from online posts such as the motivating examples given above, we adopt triplet loss as our training objective. The network is trained to distinguish encoded positive and negative pairs in terms of argument relevance. For encoding, the core model (i.e., BERT) in the Siamese architecture produces embeddings for any pair of sentences regarding their relevance.

Below, (A, P, N) refers to a triplet, where A, P, and N are respectively the anchor, positive, and negative inputs. Denoting the output embeddings from the Siamese network with shared BERT networks as z_A, z_P, and z_N, triplet loss is defined as given below. We experiment with cosine similarity, Manhattan distance, and vector concatenation for D. To prevent trivial encodings such as identical or all-zero vectors, we use α ($\alpha > 0$) as a difference margin.

$$L(A, P, N) = L(z_A, z_P, z_N) = \max(D(X'_A, X'_P) - D(X'_A, X'_N) + \alpha, 0) \quad (1)$$

For the classification task, we create an additional layer using softmax for the objective function and optimize cross-

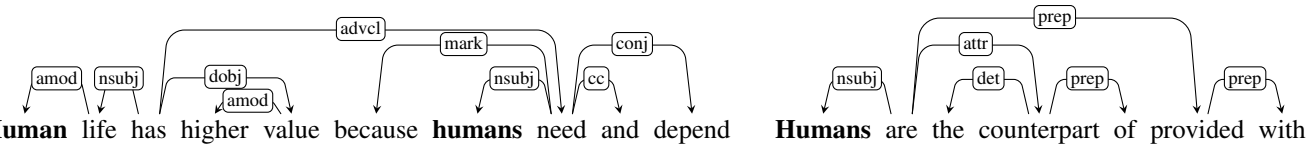

Figure 3: Dependencies between remaining words after content reduction in a pair of sentences. Take the first sentence as an instance—the connection token *human* and its one-hop word (i.e., *life*, *need*) are kept, which preserves the phrase *human life*. The root of the dependency parse tree *has*, the one-hop and two-hop words from the root (namely, *higher*, *value*, *because*, *need*, and *depend*) are kept as the "skeleton" of the sentence.

entropy loss.

$$o = \text{softmax}(\odot(u, v), |u - v|, u.v) \quad (2)$$

For generating embeddings for a pair of sentences u, v, we experiment with different computation options (represented by $\odot$) such as subtraction and concatenation ($u.v$).

Experiments and Discussion

Our experiments address two claims: (1) how effectively a model learns for capturing argumentative relevance between two sentences and (2) how effectively embeddings from a model trained for argumentative relevance benefit downstream tasks. The paired-arguments dataset and implementation of our approach will be released upon publication.

Determining Argumentative Relevance

Preprocessing, implementation, configuration. Texts from Kialo are arguments of high quality with no emojis and little social media slang, and thus require only small effort for cleaning. During preprocessing, we remove non-ASCII characters and replace any external URLs with the phrase "external link." We use the spaCy dependency parser to generate dependency trees for word-tokenized sentences. For content reduction, we use exact match on stemmed tokens from both sentences for finding overlapping words and WordNet (Miller 1995) for finding synonyms and antonyms.

In our ARM approach, we adapt the Hugging Face implementation of the sentence transformer (Reimers and Gurevych 2019a) for the shared BERT Siamese network with a modification on the triplet-loss evaluator. The statistics for our training and test sets are shown in Table 2.

Count of	P-C	A-D	R-R	Total
Training	68,749	45,352	68,749	182,850
Testing	17,289	11,396	17,289	45,974

Table 2: Statistics of pair samples for argumentative relevance classification.

We train the proposed model with triplets (A, P, N) extracted from Kialo. *P-C* pairs are used as positive samples, and *A-D* and *R-R* pairs are used as negative samples. To generate (A, P, N) triplets from pairs, we taken the first sentence in a *P-C* pair as the anchor sentence (as described in Section) and compare with the first sentence in an A-D or R-R pair and form a triplet if they are the same.

In terms of training a hidden representation (i.e., the output embeddings from the Siamese network), we use 45,352 (A, P, N) samples for training and the other 11,396 triplet samples for testing with a triplet-loss objective and cosine similarity. The evaluation metric is accuracy where the distance between an anchor sentence and a positive sentence is smaller than that between an anchor sentence and a negative sentence. On the validation set (10% held-out data), we achieve a 90.85% accuracy to identify the more relevant (i.e., the positive) sentences to anchor sentences.

In addition to triplet training and validating, to evaluate in a more general setting, we examine whether a trained ARM model can generate embeddings from text and decide argumentative relevance for a pair of sentences.

We compare the proposed approach to the following baselines: (1) a simple transformer-based encoding, BERT (Devlin et al. 2019), which maps a sentence to a 768-dimension vector with max pooling; (2) averaged word embeddings with GloVe (Pennington, Socher, and Manning 2014), which maps a sentence to a 300-dimension vector; (3) Contrastive-Tension (CT) based transformers which are pretrained for relevant tasks - Semantic Textual Similarity (STS) and Natural Language Inference (NLI) (Carlsson et al. 2021); and (4) Siamese-network-based transformer Sentence-BERT. For non-Siamese baselines, we fine-tune the models with Kialo argument pairs, and for Siamese-based Sentence-BERT (S-BERT), we fine-tune the model on top of Kialo triplets. For the fine-tuning process, we use a batch size of 32, Adam optimizer with learning rate $2\text{e}{-5}$, and 1,000 warm-up steps.

The overall results are shown in Table 3. Recall that in the original test set, P-C pairs are considered high-relevance and A-D and R-R are considered low-relevance. In addition, we created a small gold standard set. We randomly selected 1,000 pairs from the test set and labeled them. This gold standard set contains 444 low-relevance sentence pairs and 556 high-relevance pairs. We report model performance on both sets. Based on the results, Siamese-based transformers (i.e., Sentence-BERT and our approach) perform better than a nontransformer (e.g., GloVe) and a transformer in the original structure (BERT) for measuring argumentative connections. Contrastive-Tension based models do not perform as well as Siamese-based models for distinguishing argumentative relevance from topical relevance. Our approach with content reduction outperforms all other models.

We conduct statistical tests using McNemar's test to bet-

Model	Accuracy		Precision		Recall		F1-score	
	Orig	Gold	Orig	Gold	Orig	Gold	Orig	Gold
GloVe	0.5537	0.5350	0.5549	0.5392	0.5537	.5350	0.5517	0.5363
CT-NLI	0.5374	0.5730	0.5391	0.5932	0.5374	0.5730	0.5330	0.5711
CT-STSb	0.5699	0.5850	0.5734	0.6042	0.5699	0.5850	0.5651	0.5837
BERT	0.6174	0.6320	0.6190	0.6366	0.6174	0.6320	0.6160	0.6331
S-BERT	0.7344	0.7170	0.7350	0.7224	0.7340	0.7170	0.7337	0.7178
ARM	**0.8418**	**0.8060**	**0.8418**	**0.8097**	**0.8419**	**0.8060**	**0.8419**	**0.8032**

Table 3: Accuracy, precision, recall, and F1 score compared with baseline embedding methods. We report results for the original test set with automated labels (as in column Orig) and the expert-annotated gold standard set (as in column Gold).

ter understand the performance difference between the proposed model and the baseline models. We achieve p-values < 0.001 when comparing our approach to any baseline model. The odds ratio is 2.5794 with a standard error of 0.0184. In addition, we examine the effect size of our model compared to the plain Siamese model S-BERT with Cohen's g index (also known as Cohen's d_s). We achieve a score of 0.220, which indicates a medium effect size (Cohen 1988).

Ablation Study To take a closer look at the importance of the various components of our approach, we examine performance gains from different modules added on top of the Siamese architecture: the content-reduction module, the triplet-loss training objective compared to pairwise regression objective, and the usage of embeddings for classification. The results are shown in Table 4a.

Triplet loss is essential and reflects our observation that argumentative relevance (dialectical connections in addition to semantic similarity) between arguments differs from mere textual similarity between semantically similar sentences. Content reduction contributes to the results. Compared to the concatenation of hidden representations used in the optimal setup for ARM, there is not much difference between Manhattan distance or cosine similarity, both of which perform worse than concatenation.

We experiment with pretrained BERT as the core model of the Siamese architecture on classification tasks considering the following configurations: (1) use BERT (Devlin et al. 2019) or RoBERTa (Liu et al. 2019) as the core model for the Siamese network, (2) pretrain on NLI tasks or on paraphrase tasks, and (3) use mean tokens or max-pooling. We find that BERT pretrained on NLI tasks with mean tokens (Reimers and Gurevych 2019b) performs slightly better than the other models, so we report results from ARM with this configuration. Table 4b shows results from other configurations.

Applying Argumentative Relevance Mining

We now evaluate how well a trained ARM (i.e., the best-performing variant) benefits other argumentation tasks.

Predicting Impact Score. Assessment of argument quality and persuasiveness is an important challenge in computational argumentation and is a useful downstream task of our work. In online debates, argument quality assessment not only considers the soundness of an individual argument but also the timeliness and appropriateness of the argument in the context of reasoning (Durmus, Ladhak, and Cardie 2019b). *Timeliness* and *appropriateness*, in our case, are represented as logical and topical relevance of an argument to its surrounding arguments. Therefore, our first experiment on the application of ARM is to predict how impactful an argument is in an online discussion.

We evaluate the usefulness of the generated embeddings for determining argument impact in Kialo discussions (Durmus, Ladhak, and Cardie 2019b). For each argument on Kialo, users vote on its impact within the given "line" (meaning a path in Kialo discussion tree) of reasoning. We use the same data and experimental setup as Durmus, Ladhak, and Cardie (2019b): (1) we predict argument impact as one of three classes—NOT IMPACTFUL, OF MEDIUM IMPACT, and IMPACTFUL, (2) there are in total 7,386 arguments, among which 1,633 are NOT IMPACTFUL, 1,445 are OF MEDIUM IMPACT, and 4,308 are IMPACTFUL, and (3) for testing, 15% of arguments are held out.

To compare with our approach, we take the models used by Durmus, Ladhak, and Cardie (2019b) as baselines: (1) linguistic features which include TF-IDF for unigrams and bigrams, model verbs, usages of exclamation, hedging, POS n-grams, and other lexicon-based features, (2) FastText (Joulin et al. 2017), which shows competitive performance with deep learning models for text classification, and (3) fine-tuned BERT for sequence classification.

Table 5 shows the resulting precision, recall, and F1 scores. ARM outperforms previous methods in a pairwise setting (i.e., using texts from claims and their parents) with 4.55% increase in F1 score. ARM even outperforms the previous best-performing BERT model with additional preceding arguments. The notation *f3* for the reported result in Table 5 means three preceding arguments in a branch of a Kialo discussion are used as context. Linguistic features used in previous work show higher precision but much lower recall and F1. Considering identifying relevant arguments as an information mining task, a low recall score would cause less relevant information being retrieved. Therefore, transformer-based methods, especially our ARM model, perform better, which is reflected in the F1 score. Linguistic features include token (i.e., word or punctuation)-based clues such as hedge words, exclamation points, and domain-related frequent bigrams that are meaningful and indicative. For instance, "however" is a highly precise indicator for detecting contrastive relations.

Model	Acc.	Prec.	Rec.	F1
ARM (optimal)	**0.8418**	**0.8418**	**0.8419**	**0.8419**
ARM w/ regr.	0.8034	0.8013	0.8034	0.8001
ARM -CR	0.8284	0.8291	0.8284	0.8287
ARM regr. -CR	0.7519	0.7519	0.7519	0.7519
ARM w/ cos.	0.8109	0.8143	0.8109	0.8120
ARM w/ Manh.	0.8126	0.8124	0.8126	0.8125

(a) Performance with or without specific modules.

Model	Acc.	Prec.	Rec.	F1
BERT w/ paraphrase and max pooling	0.7519	0.7519	0.7519	0.7519
BERT w/ paraphrase and mean tokens	0.8079	0.8099	0.8079	0.8086
BERT w/ NLI and max pooling	0.8108	0.8336	0.8333	0.8334
RoBERTa w/ paraphrase and max pooling	0.6965	0.6876	0.6965	0.6815
RoBERTa w/ paraphrase and mean tokens	0.7340	0.7350	0.7340	0.7337
RoBERTa w/ NLI and mean tokens	0.8086	0.8120	0.8086	0.8097

(b) Performance of the proposed model with different Siamese configurations.

Table 4: Accuracy, precision, recall, and F1 score for the proposed model but with different module options.

	Precision	Recall	F1
Linguistic features	65.67	38.58	35.42
FastText	51.18	46.09	47.06
BERT (claim only)	53.24	50.93	51.53
BERT (w/ parent)	55.79	53.54	54.00
BERT (w/ context$_{f3}$)	57.19	55.77	55.98
ARM	58.11	**60.01**	**58.55**

Table 5: Results on predicting impact of arguments.

Identifying Propositional Relations. To examine the generalization of our model, we evaluate our argumentative relevance measurement to a dataset with different characteristics. In this experiment, we apply our embedding approach on another corpus—US2016reddit (Visser et al. 2018, 2019)—for an argumentation-related downstream task, namely, predicting propositional relations between two arguments. US2016 is a set of annotated corpora on dialogical argumentation. It contains transcripts of TV debates about 2016 presidential elections and online discussions from Reddit corresponding to the TV debates. Since our focus is on online discussions, we use the US2016reddit subset of the US2016 corpus that excludes the TV debate subset. The propositional relations (906 sample pairs) between two arguments are categorized as inference (56.3%), conflict (30.1%), and rephrase (13.6%). These propositional relations are categorized based on logical reasoning and dialogical communicative dimensions of two arguments in discourse analysis and argumentation theory.

For this experiment on transfer learning ability, we use the best performing model trained on Kialo, freeze its parameters, and fine-tune the last layer of classification on US2016reddit. The weighted F1 score for all three types is 0.6568, outperforming BERT-based embeddings by 8.4%. Table 6 shows a breakdown of the results by propositional type as well as challenging examples (i.e., false negatives).

Qualitative Analysis and Discussions

We discuss the benefits and limitations of the proposed model based on the experiments above.

Argumentative discussions in social media such as online debate forums (e.g., createdebate.com, debate.org), Wikipedia Talk pages, and a substantial number of rational arguments on Twitter have similar characteristics of arguments on Kialo—thus all such settings can benefit from our study. However, when applying our approach to a corpus with quite different characteristics, the results from our secondary experiments on Reddit data indicate that preprocessing and tuning on the content reduction module is needed.

The experiment on propositional relations of Reddit arguments reflects some limitations of argumentative relevance measurement. One observation is that argumentative relevance measurement may need to incorporate contexts or external knowledge to tackle challenging examples, such as those in Table 6. *Sent.A* and *Sent.B* have similar syntactic connections in both *Inference* and *Rephrase*, so the key problem is to distinguish relations between (be sniffing, have Parkinson's) and relations between (have pneumonia, be ill) to better distinguish the two propositional types. In the *Conflict* example, *Sent.B* is not necessarily in conflict with *Sent.A* without a context of participants' stances on Trump.

Our work makes a step towards identifying chains of reasoning by proposing an approach specifically for measuring argumentative relevance among online arguments. As mentioned in Section , identifying argumentatively relevant pairs from massive online posts helps form reasoning chains, which is an important dimension for argument quality, and thereby helps with evaluating online argumentative discussions and their outcomes.

ARM is geared toward relevance and need not be helpful on tasks where relevance is not the main underlying intuition. It is not well suited to stance classification, since there is no significant difference regarding logic-related syntactic structures between supporting and attacking arguments. It is also not well suited to ASPECT similarity (Reimers et al. 2019) (determining whether two sentence-level arguments are similar), which requires deep semantic representation.

Additional Related Work and Future Directions

We consider additional related work in the following themes: research in argumentation that investigates interlocutor relations, research domains that can benefit from our argumentative relevance modeling, and complementary work that may be combined with our work in future enhancements.

Little research calls attention to relations between arguments from multiple interlocutors. Chakrabarty et al. (2019) leverage Rhetorical Structure Theory for identifying argu-

	Precision	Recall	F1	*Sent.A*	*Sent.B*
Inference	0.7087	0.7526	0.7300	hillary is sniffing so much	clinton has parkinson's
Rephrase	0.5652	0.5416	0.5532	trump must have pneumonia	trump must be ill
Conflict	0.6071	0.5574	0.5812	trump has makeup on lol	pretty much anyone who appears on television has a makeup on
Overall	0.6558	0.6593	0.6568	–	–

Table 6: Results and challenging examples (false negatives) on predicting propositional relations for US2016reddit.

ment components in persuasive discussions.

In contrast to typical work on argumentation, which targets a specific problem, our work focuses on producing a general representation for relevance and can readily combine with downstream tasks. We consider various research domains, such as argument quality assessment (Skitalinskaya, Klaff, and Wachsmuth 2021), conversation derailment prediction (Yuan and Singh 2023), and argument search engine (Wachsmuth et al. 2017b), as potential downstream tasks (i.e., applicable domains). For instance, Swanson, Ecker, and Walker (2015) train a regressor to predict the quality of arguments. Gleize et al. (2019) apply Siamese architecture to assess the convincingness of evidence under certain topics. Xi and Singh (2023) incorporate language features for understanding blame assignment in online discussions. Our proposed work can be an alternative embedding method or an additional feature to consider in those works.

Argumentative relevance representation can be improved by incorporating external knowledge. Some research studies ways and benefits of leveraging information beyond text. We consider those techniques as supplementary work that can be potentially incorporated in the future. For instance, Zhao et al. (2021) leverage topic embedding and user information embedding for argument persuasion. Cui and Hershcovich (2021) investigate implicit and underspecified language in arguments. Yuan et al. (2021) leverage knowledge graph to identify related concepts between arguments, which can be a future direction to improve connection recognition.

Conclusions

The task of measuring argumentative relevance between online posts fills a gap in modeling relations between texts in terms of their argument-structure relation, especially for informal, multi-interlocutor arguments. We present a dataset of pairwise sentences with systematically generated labels regarding whether they are argumentatively relevant or not. We propose an approach based on Siamese networks with content reduction to learn a hidden representation of argumentative connections. Our approach outperforms strong baselines for determining whether a sentence is argumentatively relevant to the other sentence.

Argumentative relevance potentially applies to any argument-related tasks where both logical and topical connections between arguments are factors of interest. We demonstrate that modeling argumentative relevance benefits predicting online argument impact. The proposed method is an alternative to general embeddings which do not distinguish between semantic relevance and argumentative relevance between two sentences.

It can be generalized to produce embeddings for other information retrieval tasks. For instance, it may help with identifying interactive arguments from online platforms where users' posts are structured in a flat and loose hierarchy (e.g., Tweets with the same hashtag but not in a retweeting chain, or product reviews on Amazon). We discuss applicable scenarios and limitations and defer exploring additional applications or incorporating other techniques like knowledge graphs to future work.

Broader Impact and Ethical Use

Information overload on social media, which contains repeated, noisy, and uninformative content, reduces people's efficiency in knowledge acquisition. This work contributes to argument-mining tasks such as mining coherent arguments and assessing the completeness and effectiveness of an argumentative discussion.

Potential privacy and ethical concerns do not arise directly from this research but may from its potential applications. Although our dataset is anonymous and does not trace user-related information such as user names or historical posts, the proposed approach may help pinpoint impactful posts and influential users. Besides, for such text mining where the content is public on social media, there is a possibility that the owner of a comment may be traced by searching the content online. We have obtained proper consent from data providers and human annotators for this study.

Acknowledgments

We thank the anonymous reviewers for their comments. We thank the NCSU Laboratory for Analytic Sciences (LAS) and the US National Science Foundation (grant IIS-1908374) for partial support.

References

Cabrio, E.; and Villata, S. 2018. Five Years of Argument Mining: a Data-driven Analysis. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, 5427–5433. Stockholm, Sweden: International Joint Conferences on Artificial Intelligence Organization.

Carlsson, F.; Gyllensten, A. C.; Gogoulou, E.; Hellqvist, E. Y.; and Sahlgren, M. 2021. Semantic Re-tuning with Contrastive Tension. In *International Conference on Learning Representations*.

Chakrabarty, T.; Hidey, C.; Muresan, S.; McKeown, K.; and Hwang, A. 2019. AMPERSAND: Argument Mining for PERSuAsive oNline Discussions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2933–2943. Hong Kong, China: Association for Computational Linguistics.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Cui, R.; and Hershcovich, D. 2021. Great Service! Fine-grained Parsing of Implicit Arguments. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, 65–77. Online: Association for Computational Linguistics.

Dessalles, J.-L. 2016. A Cognitive Approach to Relevant Argument Generation. In Baldoni, M.; Baroglio, C.; Bex, F.; Grasso, F.; Green, N.; Namazi-Rad, M.-R.; Numao, M.; and Suarez, M. T., eds., *Principles and Practice of Multi-Agent Systems*, 3–15. Cham: Springer International Publishing.

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 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Dung, P. M.; and Son, T. C. 2001. An Argument-Based Approach to Reasoning with Specificity. *Artificial Intelligence*, 133(1): 35–85.

Durmus, E.; Ladhak, F.; and Cardie, C. 2019a. Determining Relative Argument Specificity and Stance for Complex Argumentative Structures. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4630–4641. Florence, Italy: Association for Computational Linguistics.

Durmus, E.; Ladhak, F.; and Cardie, C. 2019b. The Role of Pragmatic and Discourse Context in Determining Argument Impact. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5668–5678. Hong Kong, China: Association for Computational Linguistics.

Fan, Y.; Guo, J.; Ma, X.; Zhang, R.; Lan, Y.; and Cheng, X. 2021. A Linguistic Study on Relevance Modeling in Information Retrieval. In *Proceedings of the Web Conference 2021*, WWW '21, 1053–1064. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383127.

Gleize, M.; Shnarch, E.; Choshen, L.; Dankin, L.; Moshkowich, G.; Aharonov, R.; and Slonim, N. 2019. 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*, 967–976. Florence, Italy: Association for Computational Linguistics.

Guo, Z.; Zhang, Z.; and Singh, M. 2020. In Opinion Holders' Shoes: Modeling Cumulative Influence for View Change in Online Argumentation. In *Proceedings of The Web Conference 2020*, WWW '20, 2388–2399. New York, NY, USA: Association for Computing Machinery. ISBN 9781450370233.

Habernal, I.; and Gurevych, I. 2017. Argumentation Mining in User-Generated Web Discourse. *Computational Linguistics*, 43(1): 125–179.

Hidey, C.; Musi, E.; Hwang, A.; Muresan, S.; and McKeown, K. 2017. Analyzing the Semantic Types of Claims and Premises in an Online Persuasive Forum. In *Proceedings of the 4th Workshop on Argument Mining*, 11–21. Copenhagen, Denmark: Association for Computational Linguistics.

Jo, Y.; Bang, S.; Manzoor, E.; Hovy, E.; and Reed, C. 2020. Detecting Attackable Sentences in Arguments. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1–23. Online: Association for Computational Linguistics.

Jo, Y.; Poddar, S.; Jeon, B.; Shen, Q.; Rosé, C.; and Neubig, G. 2018. Attentive Interaction Model: Modeling Changes in View in Argumentation. In *Proceedings of the 16th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 103–116. New Orleans, Louisiana: Association for Computational Linguistics.

Joulin, A.; Grave, E.; Bojanowski, P.; and Mikolov, T. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 427–431. Valencia, Spain: Association for Computational Linguistics.

Lauscher, A.; Ng, L.; Napoles, C.; and Tetreault, J. 2020. Rhetoric, Logic, and Dialectic: Advancing Theory-based Argument Quality Assessment in Natural Language Processing. In *Proceedings of the 28th International Conference on Computational Linguistics*, 4563–4574. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Li, J.; Durmus, E.; and Cardie, C. 2020. Exploring the Role of Argument Structure in Online Debate Persuasion. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8905–8912. Online: Association for Computational Linguistics.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv*, abs/1907.11692.

Miller, G. A. 1995. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11): 39–41.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Persing, I.; and Ng, V. 2017. Why Can't You Convince Me? Modeling Weaknesses in Unpersuasive Arguments. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 4082–4088. Melbourne, Australia: IJCAI.

Reimers, N.; and Gurevych, I. 2019a. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. Hong Kong, China: Association for Computational Linguistics.

Reimers, N.; and Gurevych, I. 2019b. sentence-transformers/bert-base-nli-mean-tokens. Last accessed: 2021-05-14.

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.

Risch, J.; and Krestel, R. 2020. Top Comment or Flop Comment? Predicting and Explaining User Engagement in Online News Discussions. In *ICWSM*.

Schluter, N.; and Varab, D. 2018. When data permutations are pathological: The case of neural natural language inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 4935–4939. Brussels, Belgium: Association for Computational Linguistics.

Schneider, J.; Davis, B.; and Wyner, A. 2012. Dimensions of Argumentation in Social Media. In ten Teije, A.; Völker, J.; Handschuh, S.; Stuckenschmidt, H.; d'Acquin, M.; Nikolov, A.; Aussenac-Gilles, N.; and Hernandez, N., eds., *Knowledge Engineering and Knowledge Management*, 21–25. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-33876-2.

Skitalinskaya, G.; Klaff, J.; and Wachsmuth, H. 2021. Learning From Revisions: Quality Assessment of Claims in Argumentation at Scale. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 1718–1729. Online: Association for Computational Linguistics.

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

Swanson, R.; Ecker, B.; and Walker, M. 2015. Argument Mining: Extracting Arguments from Online Dialogue. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 217–226. Prague, Czech Republic: Association for Computational Linguistics.

Tan, C.; Niculae, V.; Danescu-Niculescu-Mizil, C.; and Lee, L. 2016. Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions. In *Proceedings of the 25th World Wide Web Conference*, WWW, 613–624. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee.

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

Visser, J.; Duthie, R.; Lawrence, J.; and Reed, C. 2018. Intertextual Correspondence for Integrating Corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA).

Visser, J.; Konat, B.; Duthie, R.; Koszowy, M.; Budzynska, K.; and Reed, C. 2019. Argumentation in the 2016 US Presidential Elections: Annotated corpora of television debates and social media reaction. *Language Resources and Evaluation*, 54: 123–154.

Wachsmuth, H.; Naderi, N.; Hou, Y.; Bilu, Y.; Prabhakaran, V.; Thijm, T. A.; Hirst, G.; and Stein, B. 2017a. 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*, 176–187. Valencia, Spain: Association for Computational Linguistics.

Wachsmuth, H.; Potthast, M.; Al-Khatib, K.; Ajjour, Y.; Puschmann, J.; Qu, J.; Dorsch, J.; Morari, V.; Bevendorff, J.; and Stein, B. 2017b. 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.

Walton, D.; and Reed, C. 2003. Diagramming, Argumentation Schemes and Critical Questions. In Van Eemeren, F. H.; Blair, J. A.; Willard, C. A.; and Snoeck Henkemans, A. F., eds., *Anyone Who Has a View: Theoretical Contributions to the Study of Argumentation*, 195–211. Dordrecht: Springer Netherlands. ISBN 978-94-007-1078-8.

Wang, H.; Sun, D.; and Xing, E. P. 2019. What if We Simply Swap the Two Text Fragments? A Straightforward yet Effective Way to Test the Robustness of Methods to Confounding Signals in Nature Language Inference Tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 7136–7143.

Xi, R.; and Singh, M. P. 2023. The Blame Game: Understanding Blame Assignment in Social Media. *IEEE Transactions on Computational Social Systems (TCSS)*, 10: 1–10.

Yuan, J.; and Singh, M. P. 2023. Conversation Modeling to Predict Derailment. In *Proceedings of the 17th International AAAI Conference on Web and Social Media (ICWSM)*, 1–8. Limassol, Cyprus: AAAI Press. To appear.

Yuan, J.; Wei, Z.; Zhao, D.; Zhang, Q.; and Jiang, C. 2021. Leveraging Argumentation Knowledge Graph for Interactive Argument Pair Identification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2310–2319. Online: Association for Computational Linguistics.

Zhao, X.; Durmus, E.; Zhang, H.; and Cardie, C. 2021. Leveraging Topic Relatedness for Argument Persuasion. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 4401–4407. Online: Association for Computational Linguistics.