Exploiting Dialogue Acts and Context to Identify Argumentative Relations in Online Debates

Stefano Mezza, Wayne Wobcke and Alan Blair
School of Computer Science and Engineering
UNSW Sydney, NSW 2052 Australia
{s.mezza|w.wobcke|a.blair}@unsw.edu.au

Abstract

Argumentative Relation Classification is the task of determining the relationship between two contributions in the context of an argumentative dialogue. Existing models in the literature rely on a combination of lexical features and pre-trained language models to tackle this task; while this approach is somewhat effective, it fails to take into account the importance of pragmatic features such as the illocutionary force of the argument or the structure of previous utterances in the discussion; relying solely on lexical features also produces models that over-fit their initial training set and do not scale to unseen domains. In this work, we introduce ArguNet, a new model for Argumentative Relation Classification which relies on a combination of Dialogue Acts and Dialogue Context to improve the representation of argument structures in opinionated dialogues. We show that our model achieves state-of-the-art results on the Kialo benchmark test set, and provide evidence of its robustness in an open-domain scenario.

1 Introduction

Argumentative Dialogues are discussions between two or more parties involving an opinionated topic, i.e. any topic which may divide the interlocutors into a number of conflicting opinions. These discussions are usually different from ordinary conversations, in that the speakers' goal is usually to convince their interlocutors of their own point of view by defending their own stance and attacking their opponents' arguments. Figure 1 shows an example of a debate from the Kialo online debate platform. A key aspect in the study of Argumentative Dialogues is identifying the relationship between an argument step in the discussion and preceding argument steps introduced by other speakers; this task is commonly referred to as *Argumentative Relation Classification* (Stab and Gurevych, 2014), or sometimes *Argument Polarity Prediction* (Cayrol and Lagasquie-Schiex, 2005) when it only involves a binary classification between two possible relations. In this work, we will use the term **Argumentative Relation Classification**, to avoid any confusion with similar tasks such as *Sentiment Analysis* or *Stance Classification*.

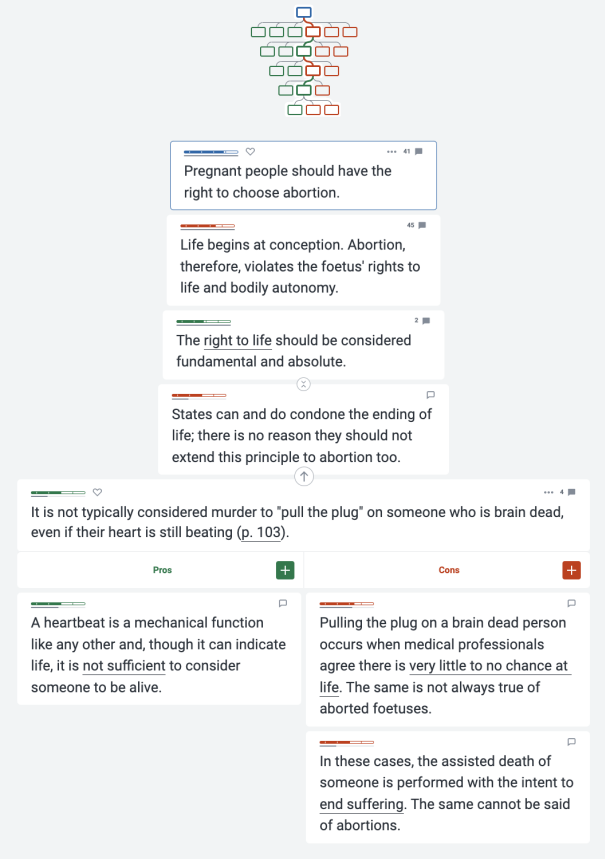

Figure 1: An example of a debate from the Kialo online debate platform. Green nodes agree with the original thesis (in blue), while red nodes disagree with it. Nodes are annotated with the *argumentative move* that they perform on their parent node in the graph (i.e. *Support* or *Attack*). Users annotate their own stance towards the thesis, as well as their argumentative move towards the node they are responding to.

Existing works in the literature that aim to

Proceedings of the 11th Workshop on Argument Mining, pages 36–45

solve this task usually rely on either hand-crafted syntactic and lexical features (Stab and Gurevych, 2014; Lenz et al., 2020), pre-trained language models (Agarwal et al., 2022; Ruiz-Dolz et al., 2021) or both (Cocarascu et al., 2020). While these models are becoming increasingly accurate, there are some shortcomings in their approach. First, they often ignore any non-lexical aspect of the dialogue, which hinders their capability to correctly understand the conversation. Second, they have limited understanding of the surrounding context of the argument contributions, and struggle to take long-term dependencies into account. Finally, they are often tested in a domain-specific scenario in which the system learns to predict relations between argument contributions that belong in the same dataset it was trained on; this makes it hard to correctly assess their capability to adapt to unseen conversations, which is crucial for practical applications such as the development of Automated Dialogue Agents.

In this work, we explore the hypothesis that contextual information and pragmatic features (such as Dialogue Act Tags) can be highly beneficial in increasing the accuracy of Argumentative Relation Classification models. We also aim at analysing how much existing models can generalise to entirely unseen topics of discussion, and how these features can help a model become less dependent on its training domain. There is evidence in the literature that Dialogue Act Tags may be used as a feature to improve a model's understanding of the argumentative structure of a debate (Petukhova et al., 2016; Budzynska et al., 2014). There is also evidence that contextual information is highly beneficial for Argument Mining tasks and, more specifically, to increase the accuracy of Argumentative Relation Classification models (Agarwal et al., 2022).

We build on this existing evidence and introduce **ArguNet**, a novel approach to Argumentative Relation Classification that relies on on a combination of Dialogue Acts and a specialised encoding of the previous nodes in the debate. ArguNet uses ISO 24617-2 Dialogue Acts (DAs) annotated with the DASHNet architecture (Mezza et al., 2022) to enrich the input utterances with additional syntactic and pragmatic information. BERT (Devlin et al., 2018) is used to encode the enriched input utterances into dense sentence embeddings, with the addition of Utterance Manipulation Strategies from Whang et al. (2021) to further increase the effectiveness of the contextual embeddings from BERT. Our approach is trained and tested on data from the Kialo online debate platform, a high-quality, publicly-available source of conversations annotated with argumentative relations. We use the same Kialo scrape introduced by Agarwal et al. (2022); however, instead of shuffling the argument contributions and dividing them in a training and test split, we split at the debate level, so that contributions from the same debate will not appear in different splits. This is done to test the hypothesis that existing models identify lexical information in the training debates and are able to use this information when tested on argument contributions from the same debates. We also sampled an additional, smaller collection of Kialo debates called *KialoAbortion* that involve discussions on reproductive rights, which we use to further test our hypothesis that Argumentative Relation Classification is highly sensitive to the topic of the classified debates.

In our experimental section, we provide evidence that the ArguNet architecture achieves state-of-the-art results on the Kialo dataset; we also provide evidence that our model outperforms existing models in the literature when tested on debates from the *KialoAbortion* test set, which shows how ArguNet can generalise to unseen domains better than existing architectures.

2 Related Work

The formal study of argumentative discussions is known in the literature as *Argumentation Theory* (van Eemeren et al., 1996). Walton (2009) divides argumentative study into four separate tasks: *identification*, which involves identifying Argumentative Dialogue Units (ADUs) in a dialogue and inserting them into a pre-determined *argumentation scheme*; *analysis*, which deals with identifying premises and conclusion of each argument; *evaluation*, which involves assessing an argument's quality and persuasive power; and *invention*, which involves the creation of novel arguments for the debate. In this work we will focus on the task of *identification* of pre-constructed ADUs in an argumentation scheme.

The identification of a logical structure for

reasoning goes back to the seminal works by Pollock (1987) and (Nute, 1988), which introduced *Defeasible Logic*, a formalism in which *conclusions* are supported by *premises* that may no longer be justified when additional premises are introduced. Dung (1995) introduced an abstract theory of *Acceptability of Arguments* in which arguments are seen as a set of logical statements, and each argument can be *accepted* or *defeated* depending on whether it clashes with other arguments. Prakken (2010) elaborated on this theory and presented a framework for structured arguments in which arguments can be supported with premises that justify their validity, and other arguments can attack the speaker's viewpoint by either attacking the argument directly, or one of its premises. Cabrio and Villata (2012) combine textual entailment and argumentation graph into a unified framework that aims at automatically detecting accepted and defeated arguments based on the entailment between them. Lenz et al. (2020) adopted this scheme in their study on Argumentative Relation Classification on the Kialo corpus, and defined *Default Inference* and *Default Conflict* relations between arguments that support and attack each other respectively. The scheme was adopted by Fabbri et al. (2021), who use Natural Language Inference models to directly compute Argumentative Relations. This approach, however, does not distinguish between the semantic problem of determining logical relations between argument steps and the pragmatic problem of determining dialogue moves in a sequence of contributions.

Rosenfeld and Kraus (2016) introduced a graph-like scheme for argumentative moves in a debate called the *Bipolar Argumentation Graph* (BAG), in which claims are represented as nodes in a weighted graph, and can be supported by other claims or *premises* that can either *Support* or *Attack* each other. As the Kialo dataset uses a graph-like structure that resembles a BAG, we will sometimes use their terminology in this work, particularly when referring to the argumentative moves between argument nodes.

Various models have been proposed in the literature for the annotation of argumentation schemes. One of the earliest examples of a formal approach to Argumentative Relation Classification is Cabrio and Villata (2012), which proposes an approach based on Textual Entailment. Naderi and Hirst (2016) uses a combination of Skip-Thought Vectors and Cosine Similarity to predict argumentative relations in parliamentary debates; their work is one of the earliest that takes advantage of pre-trained word embeddings for this task. Cocarascu and Toni (2017) propose a neural architecture based on LSTM cells to annotate a multi-topic corpus which included debates on movies, technology and politics; they formulate the problem as a three-way classification problem between the classes *Attack*, *Support* and *Neither*. Cocarascu et al. (2020) proposed a set of strong baselines for argumentative relation prediction in a dataset-independent setting, which included an attention-based model and an autoencoder. Their emphasis on dataset-independent classification is highly relevant to our work; however, they do not analyse the difference between in-domain and out-of-domain accuracy for their model and they do not provide details on how they split their data when separating training and test sets.

Recently, Agarwal et al. (2022) proposed GraphNLI, a graph-based neural architecture that uses graph walking techniques to obtain contextual information, which is then encoded with RoBERTa embeddings (Liu et al., 2019). Their model was a source of inspiration for our work, as it shares our reliance on context encoding for Argumentative Relations Classification; however, their approach does not use pragmatic features like Dialogue Acts, and it also uses weighted averaging for embeddings rather than relying on a structured approach for context encoding, which we argue is less effective when trying to capture contextual information.

The idea of adopting Dialogue Acts (DAs) as input features for Argument Mining systems has been investigated before in the literature. Fouqueré and Quatrini (2013) proposed a unified framework for argumentative analysis and inference which used DAs as part of the argumentation scheme, and used it to annotate a discussion from Prakken (2008). Budzynska et al. (2014) introduced Inference Anchoring Theory (IAT), a framework designed to model arguments via a combination of argumentative moves and the DAs associated with them. Both works utilised DA schemes that are difficult to adopt due to the scarcity of annotated data. Petukhova et al. (2016) use ISO 24617-2 DAs as part of a model designed to understand the ar-

Dimension	Communicative Function
Task	PropQuestion, SetQuestion, ChoiceQuestion, Inform, Agree, Disagree, Answer, Directive, Commissive
Social	Greeting, Goodbye, Thanking, AcceptThanking, Apology, AcceptApology
Feedback	AlloFeedback

Table 1: The DASHNet tagging scheme. Tags, also known as *Communicative Functions*, are grouped in *Semantic Dimensions* which represent different aspects of utterance functions

gumentative behaviour of participants in a debate in order to predict its outcome. This is the official standard taxonomy for DA tagging, and includes domain-independent tags across various semantic dimensions that cover different aspects of the conversation (e.g. *Social Obligations*, *Feedbacks* etc.) While their study provides useful insights on how DAs can be used to model argumentative discussions, it is limited by the use of outdated ML methods for the task and was tested on a limited number of debates. In our work, we adopt ISO 24617-2 DAs due to their flexible, multi-dimensional and domain-independent taxonomy; we rely on our previous DASHNet model from Mezza et al. (2022) which achieved state-of-the-art accuracy on various benchmark test sets for DA tagging. Table 1 illustrates the DASHNet tagging scheme.

3 Methodology

3.1 Task Definition

A debate D comprises of a set of *Argument Contributions* $D = \{A_0, , ..., A_N\}$ arranged as *nodes* in a tree structure, with contribution A_0 being the root of the tree and representing the *Thesis* (or *Topic*) of the debate, and with each contribution A_j comprising one or more sentences connected to the thesis node A_0 via a sequence of nodes $A_{j-1}, ..., A_0$, which we will refer to as the *Context* of the argument. Finally, each contribution A_j is connected to its predecessor A_{j-1} with an *Argumentative Move* $M_j \in \{Support, Attack\}$. We define **Argumentative Relation Classification** as the task of automatically identifying the argumentative move M_j characterising the relation between A_j and A_{j-1}.

3.2 Data

For this study, we chose to work with data from the Kialo online debate platform [1]. We have decided to use Kialo because it is a highly-curated platform with moderated debates and a vote system for posts, which minimizes the amount of noise, ad hominem attacks and other irrelevant information in the debate. Moreover, as the dataset is moderated, it is free of indentifiable information about individuals or offensive content. Finally, there is extensive research on many aspects of the Kialo corpus, such as the argument specificity and stance of the participants (Durmus et al., 2019) and the argumentative relevance of its conversations (Guo and Singh, 2023). Kialo debates are organised in a weighted graph-like structure: nodes in the graph represent individual, fully-formed arguments from a single participant in the debate and are called *Contributions*. Contributions are linked together with weighted edges, with the weights representing the *Argumentative Relation* between the two contributions linked by the edge. Every debate graph forms a tree-like structure, with the thesis being debated as the root node of the tree; dialogues have multiple participants, and the participants construct the tree structure collectively.

We use a scrape of Kialo introduced in Agarwal et al. (2022), which we refer to as *KialoDataset*, which is a complete scrape of the website as of January 2020. We also collected a newer scrape of the website, which we refer to as *KialoAbortion*, focusing on a specific topic; we choose *Reproductive Rights* as this is a very popular and polarising debate topic at the time of writing. We made sure that no debates from the *KialoAbortion* corpus appear in the *KialoDataset* corpus, so that the former could be used in domain studies without the risk of data leakage. The *KialoDataset* corpus contains a total of 1,470 debates and 311,238 contributions, of which 1,051 debates (231,945 contributions) are used for training, 278 debates (53,699 contributions) for testing, and the remaining ones for validation. The *KialoAbortion* corpus is significantly smaller, with a total of 40 debates (10,584 contributions), of which 27 debates (8,970 contributions) are used for training, and the remaining ones for testing. Experiments in the literature sometimes split the debates without preserving their integrity; this *Single Contribution* splitting strategy produces

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

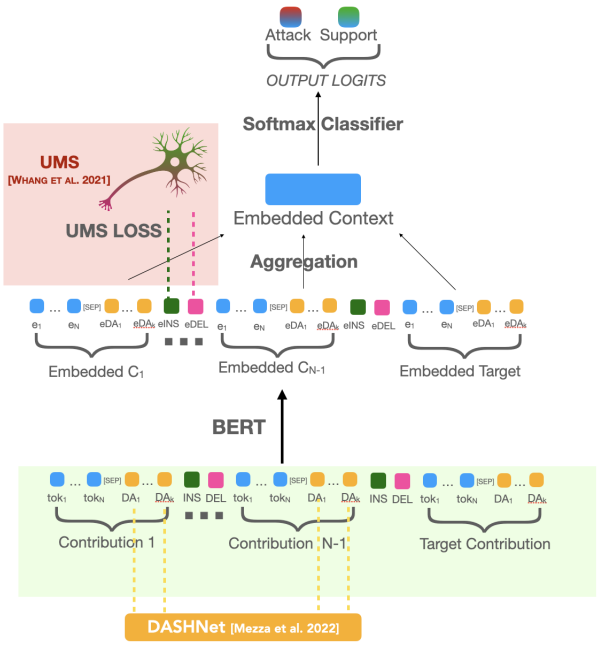

Figure 2: The ArguNet architecture.

splits which may contain argument contributions from the same debates. In contrast to that approach, we adopt a *Whole Debate* splitting strategy and split our data at the debate level, meaning that each split contains whole debates and contributions from the same debate do not appear in different splits.

3.3 Model

In this section we will outline the details of the **ArguNet** model for *Argumentative Relation Classification*. Figure 2 provides an overview of the model's architecture. ArguNet is a transformer-based architecture with a few enhancements designed to increase its accuracy when dealing with argumentative data. It uses BERT (Devlin et al., 2018) to produce dense embeddings of each token in the input arguments. In order to increase the model's ability to correctly understand each argument's underlying meaning, we enhanced the input of ArguNet with ISO 24617-2 Dialogue Act (DA) Tags extracted with the DASHNet architecture (Mezza et al., 2022). We chose the DASHNet classifier because of its multidimensional and open-domain nature, which suits our use case very well; moreover, the model uses data from the Internet Argument Corpus (Abbott et al., 2016; Walker et al., 2012), which is similar in nature and scope to the Kialo data.

ArguNet also uses Utterance Manipulation Strategies (UMS) from Whang et al. (2021) to obtain a better encoding of the context of the arguments to classify: special "[INS]" and "[DEL]" tokens are randomly inserted in the input and the corresponding utterance is either removed (in the case of "[DEL]") or erroneously inserted in the wrong spot (in the case of "[INS]"). The network has separate loss functions that control its learning of the correct UMS tags; this is combined with the classification loss from the final Softmax classifier, and the losses are averaged together to produce the final loss of the network. We added UMS to this model due to the high relevance of the order of debate turns in understanding argumentative moves; previous work acknowledged this, but leveraged context in ways that do not take into account the exact order of the utterances, such as weighted sum of embedded turns (Agarwal et al., 2022). The order of previous contributions is especially relevant to our architecture as it relies on context-aware DAs (Mezza et al., 2022); as we show in Section 4.4, UMS and DAs function especially well when combined.

Our input is an argument contribution $A_N = T_1, ..., T_M$, where T_i is the i-th token of the contribution, together with its context $C_{A_N} = A_{N-1}...A_{N-k}$, where k is the context window size of our model. We keep the window size at 5, following evidence in the literature that this is the optimal amount of context for an Argumentative Relation Classification model (Agarwal et al., 2022). We also only utilise argument contributions that directly preceded the target contribution in the debate, as opposed to alternative branches in the graph or future arguments in the discussion; this is done to make our model suitable for a real-life application in which future arguments may not be available for the analysis.

Our data is pre-annotated with the DASHNet architecture to obtain a DA-enriched argument contribution $\tilde{A_N} = T_1, ..., T_M, [SEP], DA_1, ..., DA_H$. Each contribution in the context is also annotated with its DA tags. Figure 3 shows an example of an argument contribution annotated with DASHNet Dialogue Acts; note that DASHNet tags provide structural information about single utterances in the contribution, which we argue are highly beneficial to understand Argumentative Moves in a debate. As DASHNet operates on individual utterances, a contribution may have multiple DA tags

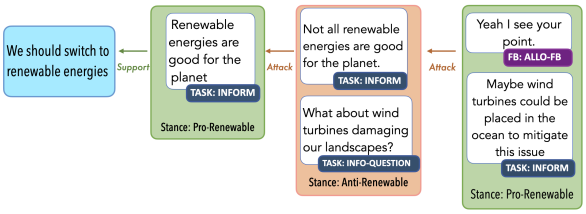

Figure 3: An example of an argument snippet annotated with Dialogue Acts. Note that DAs are annotated for each individual utterance in the contribution, and DAs might belong to different semantic dimensions (in this example, DAs from the *Task* and *Feedback* dimensions are shown)

associated with it. The input is then reshaped to utilize Utterance Manipulation Strategies, similarly to the UMS-ResSel model introduced in (Whang et al., 2021). We only utilize *Insertion* and *Deletion* strategies, as we found in our experiments that the *Search* strategies did not impact the accuracy of the resulting model when the other two strategies were present. For the insertion strategies, a target argument contribution in the context is randomly removed from its original position and placed at the end of the context window. Special [INS] tokens are placed before each contribution in the context to encode whether the target contribution should be placed in that position. Target values for the [INS] tokens are 1 for the position in which the target argument contribution originally belonged, and 0 for all other tokens. For the deletion strategies, a random outlier contribution from a different context window is randomly placed in a random place in the context. Special [DEL] tokens are placed before each argument contribution in the context to encode whether that contribution is the outlier argument or not.

The input is concatenated with its UMS-enhanced context and they are all passed to the BERT model, which produces embeddings for each token in the input (including the DA tags and the UMS tokens). A binary cross-entropy loss function is applied to the UMS tokens to determine whether the network correctly guessed the positions of the argument contributions in the context. The tokens are then stacked together to produce a dense input representation which is then fed to a Softmax Classifier similar to the one used in Sentence-BERT (Reimers and Gurevych, 2019). The final loss of the model is the sum of the classification loss and the UMS losses.

4 Experiments and Results

In this section, we illustrate the results of our experimental study. We ran three sets of experiments for this study: the first one was aimed at assessing ArguNet's accuracy when trying to determine the Argumentative Relation between two argument contributions, and compare it to existing methods in the literature, the second one was aimed at measuring the impact of each feature of the model via an ablation study, while the third one aimed at measuring how much our model and existing models rely on domain-specific lexical information in order to produce their prediction. We replicated the following models from the literature:

- **Majority Baseline:** this is just the frequency of the most prevalent argumentative move in the dataset. Both of our datasets are reasonably balanced: *KialoDataset* is comprised of 56.2% *Attack* relations and 43.8% *Support relations*, while *KialoAbortion* contains 54.8% *Attack* relations and 45.2% *Support* relations.
- **ReCAP:** this is a model trained and tested on the Kialo corpus, originally introduced in (Lenz et al., 2020) as part of a larger study on argument mining pipelines to transform textual arguments into argument graphs. The authors trained various machine learning models to predict the relation type between Kialo posts. We report results for their XG Boosting model, which is the most accurate based on our replication study.
- **BERT-Base:** this is the result of fine-tuning the BERT model on the Kialo dataset, using a single argument contribution as the context window ($k = 1$). A softmax classifier is applied to the output BERT embeddings. We chose BERT as a baseline language model since it is the foundational input embedding architecture for both ArguNet and GraphNLI.
- **GraphNLI:** this is the GraphNLI model as presented in (Agarwal et al., 2022). We used the code released by the authors, with the best-performing setting reported by the authors (weighted sum average for aggregation and *Weighted root-seeking path* with a context length of 5). As described in Section 3.2, we altered the training and test splits of the Kialo dataset to keep debates intact, rather than shuffling and splitting the argument contributions;

while we were able to replicate the authors' results with their settings, our results when evaluating this model are different from the ones they reported.

- **ArguNet:** our model (see Section 3.3).

For the ablation study, we implemented the following variations of the ArguNet model:

- **Without UMS:** this is a variation of the ArguNet model that removes the UMS from Whang et al. (2021). The DASHNet annotation is still maintained.
- **Without DA:** this is a variation of the ArguNet model that removes the DASHNet DAs. The UMS strategies are still maintained in the network.
- **Without DA and UMS**: this version of the model removes both the UMS strategies and the DASHNet annotations, leaving just the BERT embedding layer and the final Softmax Classifier.

All of the variants implemented for the ablation study maintain a context window of $k = 5$.

4.1 Implementation Details

We trained our models on Google Colab, using an NVIDIA A100 GPU with the "High RAM" setting. Training of our models took a total of roughly 400 GPU Hours, which includes all the re-trainings we had to do for our various experiments. We trained the UMS and ArguNet models for 20 epochs, but implemented early stopping with a patience of 3 (most models finished training between epochs 8 and 12). We use a Dropout rate of 0.8 for the final classification layer, a learning rate of 3e-05 and AdamW optimiser with epsilon value of 1e-8. We used BERT with 12 hidden layers, and an embedding dimension of 768, with a Dropout rate for its attention layer of 0.1. We validated all of these hyperparameters using the validation set of the *KialoDataset*. We used the "BERT base uncased" version of the BERT model from Hugging Face for any experiment involving BERT embeddings, and truncated contributions longer than 100 tokens to 100 tokens to fit the model's maximum input length of 512 (this was not generally an issue, as the average length of Kialo contributions in our data is 60 tokens. Since we had standardised training, test and validation splits for our experiments, we did not use cross-validation in our evaluation.

4.2 Argumentative Relation Classification

We trained various models from the literature on the combined train splits of the *KialoDataset* and *KialoAbortion* datasets, and compared their results to the ones obtained by the ArguNet model. We used accuracy as a metric and tested on both the *KialoDataset* and *KialoAbortion* test sets separately. All the models were trained and tested on the same data, and were trained with the *Whole Debate* splitting strategy (i.e. contributions from the same debate are kept in the same split) which produced some differences between the results we obtained and the ones reported by the authors of the respective papers. Table 2 shows the results:

Model	**Accuracy (KialoDataset)**	**Accuracy (KialoAbortion)**
Majority Baseline	54.7%	54.5%
ReCAP (Lenz et al., 2020)	66.8 %	64.1%
BERT-Base (Devlin et al., 2018)	79.2%	74.4%
GraphNLI (Agarwal et al., 2022)	79.9%	78.9%
ArguNet	**82.1%**	**81.6%**

Table 2: Argumentative Relation Classification results for our novel models, ArguNet and GraphNLI-DA, compared with other models in the literature. We replicated all models for this work.

The results show that ArguNet achieves state-of-the-art accuracy on the *KialoDataset* and *KialoAbortion* test sets. We can see that models based on BERT embeddings outperform the ReCAP model which is based on shallow machine learning methods. The GraphNLI model shows a significant decrease in accuracy on the *KialoDataset* with respect to the original result reported by the authors (82.87%): this was expected, as that result was obtained with the *Single Contribution* splitting strategy, meaning that the model would have seen other contributions from the test set during training. The model still outperforms the BERT baseline on both test sets. ArguNet shows a significant boost in accuracy over GraphNLI, which validates empirically the validity of its input encoding and context understanding strategies.

4.3 In-domain vs Out-of-domain accuracy

One of the main hypotheses that led to the design of the ArguNet architecture is that existing models in the literature largely rely on lexical information from their training corpora, which makes them less accurate when annotating debates on entirely unseen topics. In order to test this hypothesis, we compared the results of our implemented models when trained with and without the *KialoAbortion* training data. We used accuracy on the *KialoAbortion* benchmark test set as a metric. Table 3 shows the results of this study.

Model	**OOD training**	**In-domain training**	**difference (%)**
ReCAP (Lenz et al., 2020)	62.3 %	64.1%	1.8%
BERT-Base (Devlin et al., 2018)	72.3%	74.4%	2.1%
GraphNLI (Agarwal et al., 2022)	78.8%	79.9%	1.1%
ArguNet	**80.9%**	**81.6%**	**0.7**%

Table 3: Difference in accuracy between our implemented models when trained with/without in-domain data. All models were tested on *KialoAbortion*.

Results indicate that ArguNet outperforms existing approaches in the literature on both the in-domain and out-of-domain data, while also showing the lowest accuracy loss when trained without in-domain data. In general, models that utilise contextual information and other non-lexical features seem to be less prone to accuracy loss when trained without in-domain data: ReCAP and BERT-Base show significant accuracy losses (1.8% and 2.1% respectively) when trained without in-domain data, whereas GraphNLI and ArguNet exhibit much lower accuracy losses when in-domain training data is removed. This appears to validate our hypothesis that models that rely mainly on lexical features are more prone to committing annotation errors on OOD data when compared to models that adopt a more sophisticated encoding of the input.

4.4 Ablation Study

We trained various alterations of the original ArguNet architecture by removing some of its features, in order to measure their impact on the overall accuracy of the model. All variations were tested on the same test sets used in the Argumentative Relation Classification experiments. Table 4 shows the results of this study.

Model	**Accuracy (KialoDataset)**	**Accuracy (KialoAbortion)**
Without DA and UMS	79.7%	78.5%
Without DA	80.7%	80.0%
Without UMS	80.3%	79.6%
ArguNet	**82.1%**	**81.6%**

Table 4: Ablation study for the ArguNet model.

The results confirm our hypothesis that an unstructured encoding of the context is less effective than a specialised encoding, as the model trained without UMS shows a decrease in accuracy on both the *KialoDataset* and *KialoAbortion* corpora, with a 1.8% and 1.0% difference respectively. The DA feature also appears to be highly beneficial to the classification, with the "Without DA" model being significantly outperformed by the full ArguNet architecture on both the *KialoAbortion* dataset (1.4% increase) and the *KialoDataset* (1.6% increase). This follows our hypothesis that Dialogue Act Tags provide an input signal that correlates with Argumentative Relation types. The DASHNet model uses data from the Internet Argument Corpus V2 (IAC) (Abbott et al., 2016; Walker et al., 2012) ; as this corpus contains argumentative discussions that are similar in scope and style to those found in Kialo, this may also have helped the classification.

5 Conclusions

In this work, we introduced ArguNet, a neural model for the classification of Argumentative Relations between argument contributions in online debates. We showed how it achieves state-of-the-art results when tested on the Kialo dataset of online debates, and provided evidence that its defining features, namely the use of Dialogue Acts and well-structured encoding of the context of the conversation, are highly beneficial for the task at hand. Finally, we showed how its architecture is more robust to out-of-domain classification when compared to existing approaches in the literature, and provided a comparison between in-domain and out-of-domain performance for all of our baselines.

References

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

Vibhor Agarwal, Sagar Joglekar, Anthony P Young, and Nishanth Sastry. 2022. GraphNLI: A Graph-based Natural Language Inference Model for Polarity Prediction in Online Debates. In *Proceedings of the ACM Web Conference 2022*, pages 2729–2737.

Katarzyna Budzynska, Mathilde Janier, Chris Reed, Patrick Saint-Dizier, Manfred Stede, and Olena Yakorska. 2014. A Model for Processing Illocutionary Structures and Argumentation in Debates. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 917–924.

Elena Cabrio and Serena Villata. 2012. Combining Textual Entailment and Argumentation Theory for supporting Online Debates Interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 208–212.

Claudette Cayrol and Marie-Christine Lagasquie-Schiex. 2005. On the Acceptability of Arguments in Bipolar Argumentation Frameworks. In *Proceedings of the European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 378–389.

Oana Cocarascu, Elena Cabrio, Serena Villata, and Francesca Toni. 2020. Dataset Independent Baselines for Relation Prediction in Argument Mining. In *Proceedings of the 8th International Conference on Computational Models of Argument*, pages 45–52.

Oana Cocarascu and Francesca Toni. 2017. Identifying Attack and Support Argumentative Relations using Deep Learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1374–1379.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Phan Minh Dung. 1995. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and N-Person Games. *Artificial Intelligence*, 77:321–357.

Esin Durmus, Faisal Ladhak, and Claire Cardie. 2019. Determining relative argument specificity and stance for complex argumentative structures. *arXiv preprint arXiv:1906.11313*.

Alexander Richard Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. 2021. ConvoSumm: Conversation Summarization Benchmark and Improved Abstractive Summarization with Argument Mining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6866–6880.

Christophe Fouqueré and Myriam Quatrini. 2013. Argumentation and Inference: A Unified Approach. *Baltic International Yearbook of Cognition, Logic and Communication*, 8(1):4.

Zhen Guo and Munindar P Singh. 2023. Representing and determining argumentative relevance in online discussions: A general approach. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 292–302.

Mirko Lenz, Premtim Sahitaj, Sean Kallenberg, Christopher Coors, Lorik Dumani, Ralf Schenkel, and Ralph Bergmann. 2020. Towards an Argument Mining Pipeline Transforming Texts to Argument Graphs. In *Computational Models of Argument*, pages 263–270. IOS Press.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Stefano Mezza, Wayne Wobcke, and Alan Blair. 2022. A Multi-Dimensional, Cross-Domain and Hierarchy-Aware Neural Architecture for ISO-Standard Dialogue Act Tagging. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 542–552.

Nona Naderi and Graeme Hirst. 2016. Argumentation mining in parliamentary discourse. In *Principles and Practice of Multi-Agent Systems*, pages 16–25, Cham. Springer.

Donald Nute. 1988. Defeasible reasoning and decision support systems. *Decision support systems*, 4(1):97–110.

Volha Petukhova, Andrei Malchanau, and Harry Bunt. 2016. Modelling argumentative behaviour in parliamentary debates: Data collection, analysis and test case. In *Principles and Practice of Multi-Agent Systems*, pages 26–46, Cham. Springer.

John L Pollock. 1987. Defeasible reasoning. *Cognitive science*, 11(4):481–518.

Henry Prakken. 2008. A Formal Model of Adjudication Dialogues. *Artificial Intelligence and Law*, 16:305–328.

Henry Prakken. 2010. An Abstract Framework for Argumentation with Structured Arguments. *Argument & Computation*, 1(2):93–124.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Ariel Rosenfeld and Sarit Kraus. 2016. Strategical Argumentative Agent for Human Persuasion. In *Proceedings of the 22nd European Conference on Artificial Intelligence*, pages 320–328.

Ramon Ruiz-Dolz, Jose Alemany, Stella M Heras Barberá, and Ana García-Fornes. 2021. Transformer-based models for automatic identification of argument relations: A cross-domain evaluation. *IEEE Intelligent Systems*, 36(6):62–70.

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

Frans H van Eemeren, Rob Grootendorst, A Francisca Snoeck Henkemans, J Anthony Blair, Ralph H Johnson, Erik CW Krabbe, Christian Plantin, Douglas N Walton, Charles A Willard, John Woods, et al. 1996. Fundamentals of Argumentation Theory: A Handbook of Historical Backgrounds and Contemporary Developments.

Marilyn Walker, Jean E Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A Corpus for Research on Deliberation and Debate. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 812–817.

Douglas Walton. 2009. *Argumentation Theory: A Very Short Introduction*, pages 1–22. Springer, Boston, MA.

Taesun Whang, Dongyub Lee, Dongsuk Oh, Chanhee Lee, Kijong Han, Dong-hun Lee, and Saebyeok Lee. 2021. Do Response Selection Models Really Know What's Next? Utterance Manipulation Strategies for Multi-Turn Response Selection. In *Proceedings of the AAAI Conference on Artificial Intelligence, 35*, pages 14041–14049.