Leveraging Graph Structural Knowledge to Improve Argument Relation Prediction in Political Debates

Deborah Dore[1], Stefano Faralli[2], Serena Villata[1]
[1]Université Côte d'Azur, CNRS, INRIA, I3S, France
[2]Sapienza University of Rome, Italy
{deborah.dore, serena.villata}@univ-cotedazur.fr
stefano.faralli@uniroma1.it

Abstract

Argument Mining (AM) aims at detecting argumentation structures (i.e., premises and claims linked by attack and support relations) in text. A natural application domain is political debates, where uncovering the hidden dynamics of a politician's argumentation strategies can help the public to identify fallacious and propagandist arguments. Despite the few approaches proposed in the literature to apply AM to political debates, this application scenario is still challenging, and, more precisely, concerning the task of predicting the relation holding between two argument components. Most of AM relation prediction approaches only consider the textual content of the argument component to identify and classify the argumentative relation holding among them (i.e., support, attack), and they mostly ignore the structural knowledge that arises from the overall argumentation graph. In this paper, we propose to address the relation prediction task in AM by combining the structural knowledge provided by a Knowledge Graph Embedding Model with the contextual knowledge provided by a fine-tuned Language Model. Our experimental setting is grounded on a standard AM benchmark of televised political debates of the US presidential campaigns from 1960 to 2020. Our extensive experimental setting demonstrates that integrating these two distinct forms of knowledge (i.e., the textual content of the argument component and the structural knowledge of the argumentation graph) leads to novel pathways that outperform existing approaches in the literature on this benchmark and enhance the accuracy of the predictions.

1 Introduction

Argument Mining (AM) is the subfield of Natural Language Processing (NLP) that deals with automatically extracting argument structures (e.g., premises, claims, support and attack relations) from text (Lawrence and Reed, 2019; Arora et al., 2023). Argumentation graphs are then built where the identified argument components are the nodes of the graph and the edges represent support and attack relations among the components. Extracting argument structures has key applications in political scenarios (Menini et al., 2018; Visser et al., 2020a; Goffredo et al., 2022; Mancini et al., 2022) as making explicit the underlying argumentation graph of a political debate can unveil underlying strategies, inconsistencies, persuasive tactics and logical fallacies in the arguer's statements.

AM includes two main sub-tasks: *(i)* the identification of argument components, such as *claims* and *premises*, and their boundaries; *(ii)* the prediction of the relation, e.g., *support* or *attack*, holding between these components. In literature, different approaches showed promising results on the two tasks (Lippi and Torroni, 2016; Niculae et al., 2017; Stab and Gurevych, 2017; Mayer et al., 2021; Morio et al., 2022; Mushtaq and Cabessa, 2023).

The performance of AM models deteriorates when applied on political debates (Ruiz-Dolz et al., 2021; Goffredo et al., 2023b), given the complexity of the argumentation proposed in this context. The task of relation prediction, particularly when applied to political debates, has proven to be particularly challenging due to the small number of manually annotated resources for this task (Haddadan et al., 2019b; Visser et al., 2020a,b) and the lack of standard baselines against which to compare (Gemechu et al., 2024). Most existing methods in the literature predict the relations between argument components based solely on the textual content of the argument, ignoring the structure of the whole argumentation graph and the connections of the involved premises and claims towards other argument components in the graph. To address this challenging issue, recent approaches proposed frameworks that incorporate structural knowledge to achieve better results in the AM task (Khatib et al., 2020; Yuan et al., 2021). Their results are

Proceedings of the 12th Argument Mining Workshop, pages 74–86
July 31, 2025 ©2025 Association for Computational Linguistics

highly encouraging, providing even stronger support for leveraging technologies that combine structural knowledge with AM techniques. In this paper, we answer the following research questions:

RQ1: Can structural knowledge contained in Knowledge Graphs be profitably employed in challenging tasks such as argument relation prediction?

RQ2: If so, can we integrate Knowledge Graph models with existing AM models to improve the state-of-the-art (SOTA) on the argument relation prediction task?

Our proposal consists in taking a different perspective on the argument relation prediction task, by integrating the structural information of the underlying argumentation graph into the classification task. We evaluated our novel approach on a standard challenging benchmark in the AM field for political debates, i.e., the *ElecDeb60to20* dataset (Goffredo et al., 2023b). This dataset is, to the best of our knowledge, the largest available dataset of political debates manually annotated with argument components and relations.

More precisely, our approach leverages structural knowledge in the form of a Knowledge Graph (KG), i.e., a structured representation of facts through entities, relationships, and semantic descriptions. Entities represent either word objects or abstract concepts, while relations represent the connections between entities. To leverage the knowledge contained in the KG, we employ Knowledge Graph Embedding Models (KGEMs) (Bordes et al., 2013; Yang et al., 2015; Dettmers et al., 2018; Wang et al., 2021a), which are models designed to efficiently capture the semantics and the structure of a KG by mapping its entities and relations to a lower-dimensional vector space. The best-performing KGEM is integrated with a fine-tuned Language Model (LM) to improve the predictions on the argument relation classification task using a Machine Learning (ML) classifier.

The main contributions of our work are summarized as follows:

- We combine KGEMs with SOTA models in AM, leveraging fine-tuned LMs to improve SOTA results on the argument relation prediction task.
- We perform extensive experiments over several KGEMs to reveal the structural information contained in argumentation graphs.

Our hybrid approach, in its best-performing configuration, achieves a 0.73 Macro F1-Score for the argument relation prediction task, outperforming SOTA approaches on the challenging standard benchmark *ElecDeb60to20* (Haddadan et al., 2019a; Goffredo et al., 2023b). Our results show the importance of strategies that take into account structural information when dealing with NLP tasks over graph-based information, such as argument-based debates.

Furthermore, our method does not depend on joint training or new complex models, as previous approaches in the literature (Li et al., 2021; Saadat-Yazdi et al., 2023), and it represents a resource-efficient approach building on KG-based models.

The rest of the paper is organized as follows: Section 2 discusses the related work, while Section 3 illustrates the methods and the experimental setting. Section 4 and Section 5 present our findings and the error analysis. Section 6 summarizes the key outcomes.

2 Related Work

In more recent developments, pre-trained transformers like BERT have been increasingly adopted for tasks such as argument recognition, relation prediction, and premise/conclusion identification within political debates. These models leverage their deep contextual understanding to achieve significant improvements over earlier methods (Poudyal et al., 2020; Ruiz-Dolz et al., 2021).

The behavior of transformer-based models in predicting argument relations has been investigated in multiple approaches in the literature. In (Ruiz-Dolz et al., 2021), the authors applied various transformer-based models, including BERT, XL-Net, RoBERTa, DistilBERT, and ALBERT, to classify four types of relations in the IAT labeling schema: inference (RA), conflict (CA), rephrase (MA), and no relation. Their approach achieved a macro F1-score of 0.70 on the 2016 US Political Debates dataset (US2016). More recently, multimodal AM techniques have gained attention. A study on the 2020 US Political Debates (US2020) explored the integration of audio and transcript features to improve AM tasks (Mestre et al., 2021). The study on the M-Arg multi-modal dataset found that audio-only and multi-modal models performed with high accuracy and F1 scores in the argument relation classification task; However, the classification of support and attack relations remains challenging, with the highest F1 scores reaching only

0.24 and 0.21, respectively.

While initial approaches overlooked the importance of structural information, recent research underscores its critical role (Yuan et al., 2021; Morio et al., 2022). Structural knowledge—such as the relationships between different components of an argument—plays a crucial role in understanding the connections within arguments. Studies demonstrated that constructing an argumentation knowledge graph supports complex tasks like argument synthesis and question answering (Khatib et al., 2020). Their approach integrates various sources of information to enrich argument analysis.

Further innovations include the use of KGs to facilitate reasoning through argumentation paths. Graph Convolutional Networks (GCNs) have been employed to learn concept representations within KGs, coupled with a transformer-based encoder to model the paths between concepts (Yuan et al., 2021). Following this research line, some recent approach introduced the use of a Commonsense Transformer (COMET) to find inference chains connecting argumentative units (Saadat-Yazdi et al., 2023). Their proposed algorithm, ARGCON, dynamically generates these chains using the commonsense knowledge encoded in COMET, offering a novel approach to understanding argumentation. Another related study developed a topic-specialized KG by extracting evidence and identifying arguments at the sentence level (Li et al., 2021). Their hybrid model integrates topic modeling with latent Dirichlet allocation (LDA) and word embeddings to leverage both structured and unstructured data. Gemechu et al. (Gemechu and Reed, 2019) propose to combine structural and distributional techniques to achieve robust, domain-independent performance in the relation prediction task. Their model was tested on various datasets, including the US2016G1tv corpus, where it achieved an F-score of 0.64 in the classification of relations within political debates.

3 Methodology

In this Section, we detail our methodology and experimental setting. The dataset we used for this work is presented in Section 3.1, and the KGs generated from this dataset are shown in Section 3.2. The tested KGEMs are described in Section 3.3, and the tasks and metrics used to assess the models can be found in Section 3.4.

3.1 Dataset

The *ElecDeb60to20* dataset (Goffredo et al., 2023b) used in our experiments is a collection of televised political debates in the US from 1960 to 2020. The dataset consists of 44 debates featuring 64 speakers. It has been annotated with the two basic argument components - *claim* and *premise* - and with argument relations such as *support* (positive relation), *attack* (negative relation) and *equivalent* (rephrasing or restatement) (Cabrio and Villata, 2018).

The dataset comprises 38,667 argument components linked 26,230 times using the previously described relations. Among the arguments, 25078 are classified as claims, while 13589 are identified as premises. Regarding the relations, 21689 are annotated as support, 3835 as attack, and 706 as equivalent. There is a visible imbalance in the dataset: the claims are higher than the number of premises due to the tendency of candidates to make claims during political speeches without providing the necessary facts to support them (Haddadan et al., 2019c). Furthermore, the *support* relation is dominant between the relations and the *equivalent* relation is severely under-represented. For this reason, previous studies on this dataset (Goffredo et al., 2022) ignored the *equivalent* relation. The dataset mainly consists of isolated argumentation subgraphs, reflecting the debates' structure. The moderator introduces a topic (e.g., minimum wage), allows discussion, and then shifts to a new topic (e.g., relations with Cuba), repeating this process.

For training, the dataset split was 80% for training, 10% for validation and 10% for testing.

3.2 Knowledge Graph Generation

In order for the KGEMs to handle the dataset, each debate was transformed into a series of triples *(h, r, t)* where the head entity *h* and tail entity *t* represents argument components, either claims or premises, and *r* represents the relation of *support*, *attack* or *equivalence* between those components[1].

In addition to the arguments and their types (claim or premise), the dataset included information about the speaker and the year of the argument. We integrated this data and created various KG combinations, each containing different types of information. Different ad hoc relations were created to connect these additional nodes to the graph: we created the relations *says*, *year*, and *type* to connect

[1]Typically, a premise supports a claim, with *h* as the premise and *t* as the claim *t*. However, a claim can also serve as a premise to support another claim.

Ref.	Dataset	#nodes	#edges	%support	%attack	%equivalent	%type	%speaker	%year
(i)	basic	29,791	26,100	80%	15%	5%	-	-	-
(ii)	+ year node	29,835	56,064	38%	7%	1%	-	-	54%
(iii)	+ speaker node	29,855	57,868	37%	7%	1%	-	55%	-
(iv)	+ type node	29,793	63,227	34%	6%	1%	59%	-	-
(v)	+ type and year nodes	29,837	93,191	23%	4%	1%	40%	-	32%
(vi)	+ type and speaker nodes	29,857	94,995	23%	4%	1%	39%	33%	-
(vii)	+ year and speaker nodes	29,899	87,832	25%	4%	1%	-	36%	34%
(viii)	+ type, year and speaker nodes	29,901	124,959	17%	3%	0.5%	30%	25.5%	24%

Table 1: Statistics for different KG permutations. Each row represents a unique permutation incorporating various nodes and their effects on graph structure.

Ref.	Dataset	#nodes	#edges	%support	%attack	%equivalent	%type	%speaker	%year
(ix)	modified argument nodes	37,127	26,103	83%	15%	2%	-	-	-
(x)	+ speaker node	37,191	64,787	33%	6%	1%	-	60%	-
(xi)	+ year node	37,171	63,425	34%	6%	1%	-	-	59%
(xii)	+ speaker and year nodes	37,235	102,109	21%	4%	1%	-	38%	36%

Table 2: Statistics for KG permutations with dual-role argument nodes (claim and premise), including node, edge, and relation distributions.

the speaker, year, and type nodes to the appropriate argument nodes. We believe that this inclusion will increase the graph's size, decrease the number of isolated clusters, and ultimately improve the models' performance (see Table 1 for details).

As mentioned earlier, an argument can function as both a claim and a premise, depending on the context. Instead of creating a single node with the type information and linking it to the argument node via a new relationship, we explored an alternative approach: generating two separate nodes for each argument—one representing its role as a premise and the other as a claim. To differentiate these nodes, we constructed their labels by concatenating the argument text with its corresponding type. For example, the argument *It's what we are* can serve as a claim or as a premise. Therefore, we generate two distinct entities: *It's what we are_claim* and *It's what we are_premise*. We argue that this new strategy reflects the dynamic nature of arguments, where their role changes according to their relationships with other arguments. We expect that this improved representation will enhance the model's capacity to handle context-dependent argument roles (see Table 2).

To improve the models' prediction (Drance et al., 2023), we provided sentence embeddings built with Sentence-Bert (SBERT) (Reimers and Gurevych, 2019) as a starting point for the entities of the argument nodes (the only nodes containing sentences). SBERT, a refined version of BERT, is capable of producing embeddings that capture the semantic relationships within and between sentences, providing a robust foundation for representing arguments.

3.3 Knowledge Graph Embedding Models

We used three KGEMs from different categories.

TransE (Bordes et al., 2013) (translational): represents entities and relations in a continuous vector space, translating a head entity by a relation to approximate the tail entity; *DistMult* (Yang et al., 2015) (semantic matching): uses a bi-linear function to score triples, with each relation interacting multiplicatively with the embeddings of its entities; *ConvE* (Dettmers et al., 2018) (neural network based): employs Convolutional Neural Network (CNN) to model complex relationships and extract semantic information from the KG. In order to choose the KGEM and KG permutation that will best serve our goal, each KGEM is thoroughly assessed on several tasks on each KG's permutation.

3.4 Tasks & Evaluation metrics

We evaluated the KGEM on all permutation of the KG (Table 1 and 2) in different tasks (Wang et al., 2021a; Yan et al., 2022). *Link prediction* involves predicting the missing head h or tail t entity in a triple $(?, r, t)$ or $(h, r, ?)$. A variant, *relation prediction*, focuses on predicting the missing relation r in a triple $(h, ?, t)$. During evaluation, each test triple (h, r, t) is perturbed by replacing the head h with every other entity $\hat{h}$, and the resulting triples are ranked based on their scores. The goal is to rank the original triple highest. The same process

applies for predicting t and r. *Link deletion* revolves around identifying triples with erroneous head entities ($\hat{h}$, r, t) or inaccurate tail entities (h, r, $\hat{t}$). *Triple classification* and *relation classification* are the task of determining whether a triple is true (plausible) or false based on a given threshold. For *triple classification*, the evaluation protocol uses a dataset composed of 50% original triples and 50% corrupted triples, created by randomly permuting the head h, tail t, and relation r. For *relation classification*, the evaluation uses a dataset containing all original testing triples along with two permutations of each triple's relation r with incorrect relations. For both tasks, each triple's score is compared against the predefined threshold: if the score exceeds the threshold, the triple is classified as true; otherwise, it is classified as false.

To determine the predefined threshold, we calculate the median of the scores of a test dataset having 50% noise (i.e., a dataset containing 50% corrupted triples) and the median of the scores of a noise-free test dataset. Let ν represent the test dataset with 50% corrupted triples and r represent the noise-free test dataset. The threshold is calculated using Equation 1 (Faralli et al., 2023).

$$\text{threshold} = \text{median}(\nu) + \frac{\text{median}(\nu) + \text{median}(r)}{2} \quad (1)$$

Hits@k, $k \in \{1, 3, 5, 10\}$, *Mean Rank* (MR), and *Mean Reciprocal Rank* (MRR) are used to evaluate link prediction, relation prediction, and link deletion (Cao et al., 2022). Triple and relation classification, a binary classification tasks, were evaluated using *Accuracy*, *F1-Score*, *Macro F1-Score*, and *Positive and Negative F1-Score* (Powers, 2011).

3.5 Implementation Details

All experiments on the KGEMs were conducted using PyKEEN 1.8.0 (Ali et al., 2021) on Python 3.8 with an Nvidia V100 32GB GPU. For the combined architecture we also used the Hugging Face Transformers (Wolf et al., 2019) and the scikit-learn library (Pedregosa et al., 2011). We release dataset and code: `https://github.com/deborahdore/political-debates-graph-analysis`.

4 Evaluating KGEMs for relation prediction on argumentation graphs

In this Section, we answer to **RQ1**, showing how KGEMs can be successfully employed in the challenging task of relation prediction for argumentation graphs. Our benchmark is composed of two parts: in each we evaluated TransE, DistMult and ConvE (Section 3.3) using link prediction, link deletion and triple classification (Section 3.4) on all permutation of the KG (Section 3.1).

First benchmark. The first part involved the evaluation of all kinds of triples, including the one containing information related to the speaker, year and type of argument. As a random baseline we tested the model on a random composition of the KG, consisting of 50% erroneous triples and 50% correct triples for each permutation. Table 3 reports the result of link prediction, link deletion and triple classification on KG permutation setting *(i)*. All KGEMs where hyper-tuned using the default search grid of the PYKEEN library (Ali et al., 2021). The random baselines were constructed using the default hyper-parameters of the library. The study documented in Table 3 shows that the results are similar to the baseline and, in some cases, poorer.

Ref.	Model	↑ Link Prediction Hits@10	↑ Link Deletion Hits@10	↑ Triple Classification Macro F1
(i)	TransE	**0.095**	0.004	0.489
	Baseline	*0.038*	*0.004*	*0.643*
	DistMult	0.056	**0.005**	**0.494**
	Baseline	*0.011*	*0.005*	*0.526*
	ConvE	0.008	0.001	0.327
	Baseline	*0.0007*	*0.002*	*0.401*

Table 3: Benchmark results for link prediction, deletion, and triple classification tasks, compared to a random baseline on setting *(i)* of the KG..

We hypothesize that the large number of isolated components makes it difficult to correctly train the KGEMs. Interestingly, in certain cases, the random baseline generates more interconnected graphs than the original, leading to improved KGEM performance. Figure 1 demonstrate that adding connections in the KG positively impacts the performance of the KGEMs for some configuration in the triple classification task with respect to the basic KG *(i)*.

Second benchmark. During the second part of our benchmark, our evaluation will be directed toward those triples (h, r, t) whose relation r falls under *support*, *attack* or *equivalent* while still training the model with all kinds of triples.

Table 4 assesses TransE, DistMult, and ConvE using only triples of interest throughout the eval-

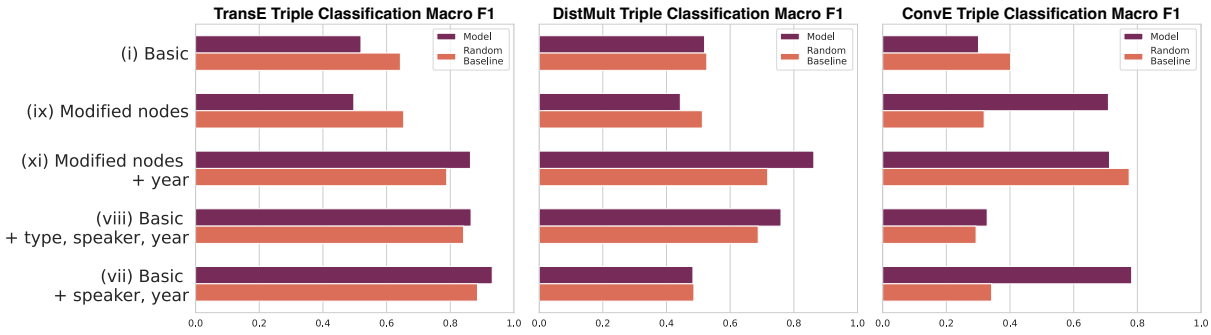


Figure 1: Comparison of the Macro F1-Score for triple classification across five KG permutations using TransE, DistMult, and ConvE.

uation, with and without pretrained embeddings built using SBERT, for the basic permutation of the KG *(i)*.

Ref.	Model	↑ Link Prediction Hits@10	↑ Link Deletion Hits@10	↑ Triple Classification Macro F1
(i)	TransE	**0.089**	0.004	**0.610**
	Baseline	*0.003*	*0.004*	*0.656*
	DistMult	0.009	**0.005**	0.283
	Baseline	*0.040*	*0.003*	*0.523*
	ConvE	0.026	0.004	0.402
	Baseline	*0.001*	*0.001*	*0.433*
(i) with pre-trained embeddings	TransE	**0.038**	**0.006**	**0.658**
	Baseline	*0.027*	*0.004*	*0.604*
	DistMult	0.017	0.002	0.509
	Baseline	*0.007*	*0.002*	*0.544*
	ConvE	0.0003	0.004	0.570
	Baseline	*0.0003*	*0.001*	*0.424*

Table 4: Performance of KGEMs on argument-specific triples with and without pre-trained embeddings, compared to random baselines on KG setting *(i)*.

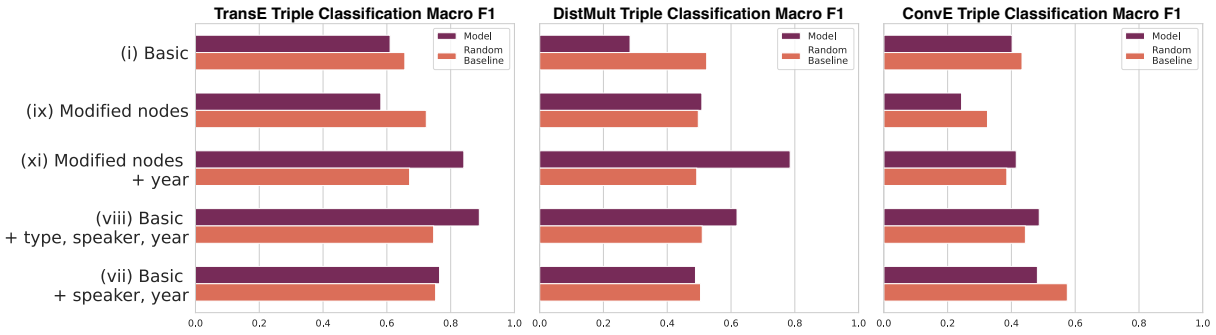
Based on our observations, the performance levels are lower when evaluating only triples of interest compared to all triples. This discrepancy is due to models focusing their attention across various types of triples, causing an incomplete evaluation of the specific triples of interest and a subsequent drop in performance.

Figures 2a and 2b present the results of the triple classification task, before and after using pre-trained embeddings, respectively. The architectural differences among KGEMs can be the reason for their diverse performances. DistMult and ConvE, with their more intricate architectures, seem to make good use of pre-trained embeddings, which enables them to identify subtle connections in complex political debates. On the other side, TransE's more straightforward design might find it difficult to make the most of the enriched embeddings, which could lead to an oversimplification of the complex relationships found in argumentation graphs from political debates.

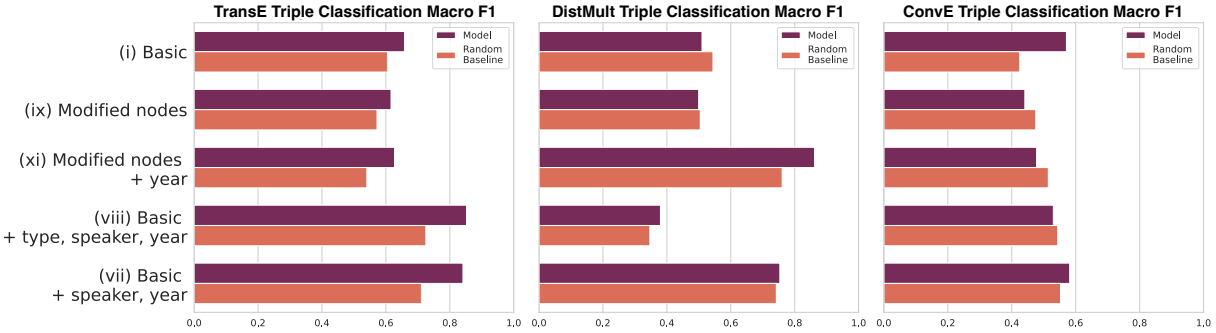
Our results indicate that KGEMs perform variably depending on the dataset and on the task: in tasks involving all types of triples the models generally performed at or above baseline levels, particularly when graph connectivity improved. This suggests that the models can capture complex relationships when the graph provides sufficient structural information. However, when we concentrated only on argumentation-specific relations such as support, attack, and equivalent, performance fell. The use of pre-trained embeddings (e.g., SBERT) improved the performance of some models, such as DistMult and ConvE, in these focused tasks. This shows how models can better represent relational dynamics in argumentation graphs by incorporating semantic enrichment from outside sources.

Despite these challenges, the models' ability to outperform random baselines in a range of configurations, as well as their improvement with more structured and enriched data, indicate that KGEMs are a viable tool for reasoning over political argumentation graphs. However, their applicability in this domain may necessitate accurate preprocessing, such as improving network connectivity or adding additional semantic data.

Error Analysis. To analyse the recurrent misclassifications, we chose the three most promising configurations with pre-trained embeddings: *(xi)*, *(viii)*, and *(vii)* (see Tables 1 and 2). Those configurations were chosen due to having the highest score average in all task among all three models. During the error analysis, the models were evaluated on tasks closer to the AM domain such as *relation prediction and classification*. Our goal was to assess each model's ability to predict and classify



(a) Evaluation of the triple classification focusing on argumentation-related triples (support, attack, equivalent).



(b) Evaluation of the triple classification of argumentation-specific triples incorporating pre-trained embeddings during training.

Figure 2: Comparison of triple classification performance across different KG configurations and the effect of pre-trained embeddings. The figure shows Macro F1-Scores for TransE, DistMult, and ConvE.

relations individually. Based on the analysis presented in Table 5, DistMult was selected as the best model for the next part of the work due to its more balanced performance across various tasks and settings.

Ref.	Model	↑ Relation Prediction Hits@1	↑ Support Prediction Hits@1	↑ Attack Prediction Hits@1	↑ Equivalent Prediction Hits@1	↑ Relation Classification Macro F1
	TransE	0.605	0.652	0.435	0.099	0.685
(xi)	DistMult	0.715	0.827	0.206	0.070	0.740
	ConvE	0.780	0.940	0.000	0.127	0.504
	TransE	0.749	0.823	0.453	0.113	0.649
(viii)	DistMult	0.153	0.010	0.990	0.000	0.366
	ConvE	0.149	0.003	0.997	0.014	0.599
	TransE	0.747	0.838	0.341	0.155	0.657
(vii)	DistMult	0.660	0.775	0.122	0.056	0.595
	ConvE	0.259	0.165	0.836	0.014	0.615

Table 5: Analysis of biases in predicting argumentation relations (support, attack, equivalent) using TransE, DistMult, and ConvE.

Although ConvE performed well in predicting argument relations, especially for the *attack* relation, it exhibited a significant bias by completely ignoring this relation in certain cases. Additionally, ConvE showed inconsistent results in triple classification, with its Macro F1-Score averaging around 50%, which indicated a lack of robustness in this task. TransE, while consistent in its predictions, suffered from skewed results due to dataset imbalance, especially in the classification of the *equivalent* relation. This made its overall performance less reliable compared to DistMult. DistMult, on the other hand, showed a more balanced performance across the different settings of the KG. It performed particularly well in setting *(xi)*, achieving the highest relation classification F1-Macro score among all models. Its performance in *(xi)* demonstrated its ability to handle the dataset's complexity effectively, making it the preferred model for the next phase of the work.

5 Integrating KGEMs with LMs to enhance relation prediction

In order to address **RQ2** (i.e., how to integrate KGEMs on existing AM models to improve the SOTA on the argument relation prediction task), we merged the tasks of relation classification and prediction. DistMult achieved a Macro F1-Score of 60%, with a precision of 66% and a recall of 60%. Previous research (Haddadan et al., 2019a) identified RoBERTa (Liu et al., 2019) as the highest performing LLM for the argument relation prediction task on the *ElecDeb60to20* dataset with a 60% Macro F1-Score. To integrate the DistMult and RoBERTa (Goffredo et al., 2023a) models, we tested different approaches, such as weighting the

predictions of DistMult and RoBERTa based on their respective Macro F1-Scores and employing a classifier to combine DistMult and RoBERTa's outputs. In this last approach, DistMult and RoBERTa are integrated using a classifier, which receives as input two features containing the prior models' predictions and returns a final prediction, as visualized in Figure 3. All tested classifiers were hypertuned using the scikit-learn library (Pedregosa et al., 2011) using the basic grid search approach.

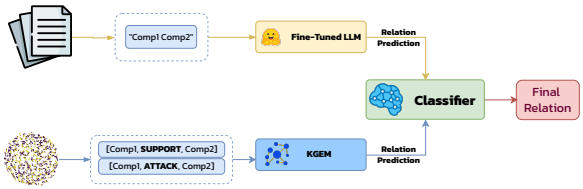


Figure 3: Proposed framework combining the LLM and KGEM, integrating predictions via a binary classifier to determine argument relations.

During inference, the arguments whose relation is to be predicted are given as input individually to the LLM and the KGEM. The LLM receives the concatenation of the arguments, *Component 1* and *Component 2*, and outputs the most likely relation (either *support*, *attack* or *no-relation* if it determines there is no relation). On the other hand, because the KGEM scores triples, it is given two triples: one with relation *support* and one with relation *attack*[2]. The triple with the highest score above the threshold is chosen as the proper one. If no triples exceed the threshold, the no-relation label is passed to the classifier.

The classifier is a machine learning (ML) model that has been trained to distinguish between the right predictions of the LLM and KGEM, as well as those that are incorrect. It returns a final relation. We selected various ML models and we evaluated them using cross-validation on a dataset composed of the predictions of DistMult and RoBERTa on their original dev and test set.

According to the findings, combining both models resulted in a 8% improvement in the state of the art for the prediction of relations between arguments in political debates using the *ElecDeb60to20* dataset. The best performing classifier is a Random Forest Classifier (RFC) (see Table 6).

We evaluated our approach using other LMs:

[2]The *equivalent* relation is not included because prior work excluded it due to its under representation in the dataset. We adopted the same approach when integrating our method into the architecture.

Integration Method	Input Type	↑ Macro F1
Random Forest	Predictions from RoBERTa and DistMult	**0.683**
AdaBoost	Predictions from RoBERTa and DistMult	**0.683**
Gradient Boosting	Predictions from RoBERTa and DistMult	**0.683**
Decision Tree	Predictions from RoBERTa and DistMult	0.680
MultiLayer Perceptron	Predictions from RoBERTa and DistMult	0.677
Support Vector Machine	Predictions from RoBERTa and DistMult	0.653
Average of Models based on their F1-Macro Score	NA	0.649
K-Nearest Neighbors	Predictions from RoBERTa and DistMult	0.642
Convolutional Neural Network	Concatenated arguments and predictions from RoBERTa and DistMult	0.639
DistMult (Single Model)	Two triples *(h,r,t)* with r ∈ support, attack	0.604
RoBERTa (Single Model)	Concatenated arguments	0.603
Gaussian Naive Bayes	Predictions from RoBERTa and DistMult	0.573

Table 6: Comparison of classifiers integrating RoBERTa and DistMult predictions for argument relation classification.

DeBERTa-V3 (He et al., 2021), BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2019) and XLM-RoBERTa (Conneau et al., 2020). DeBERTa-V3 emerged as the best model for the relation prediction task, surpassing RoBERTa with a Macro F1-Score of 69% in classifying relations between argument components (see Appendix A).

Following the same approach used with RoBERTa, we combined DeBERTa-V3 with DistMult using a classifier. The highest-performing classifier was a Convolutional Neural Network (CNN). In this case, the classifier received three input features: the predictions from DeBERTa-V3 and DistMult, and the concatenated head h and tail t arguments. This new combination achieved a 73% Macro F1-Score (see Table 7). This represents a 13% improvement over DistMult alone and a 4% improvement over DeBERTa-v3. Further analysis shows that DistMult and DeBERTa align well, predicting the same relations in 69.68% of cases. When all three models—DistMult, DeBERTa, and the classifier—agree, the prediction is correct 69.63% of the time. The classifier disagrees more often with the transformer model (16%) than with the KGEM (14%), while simultaneous disagreement with both occurs in only 0.06% of cases.

Figure 4 shows that both DistMult and DeBERTa-V3 excel at predicting the absence of a relation (*no relation*), with DistMult performing best for this class. However, both models often mis-

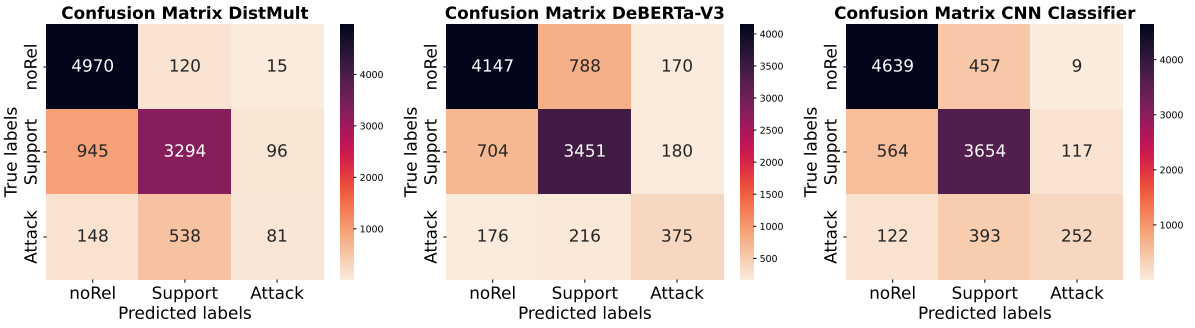


Figure 4: Confusion Matrix of DistMult, DeBERTa-V3 and the CNN classifier.

Integration Method	Input Type	↑ Macro F1
Convolutional Neural Network	Concatenated arguments and predictions from DeBERTa-v3 and DistMult	**0.734**
Average of Models based on their F1-Macro Score	NA	0.709
DeBERTa-v3 (Single Model)	Concatenated arguments	0.694
Support Vector Machine	Predictions from DeBERTa-v3 and DistMult	0.665
AdaBoost	Predictions from DeBERTa-v3 and DistMult	0.652
Gaussian Naive Bayes	Predictions from DeBERTa-v3 and DistMult	0.608
DistMult (Single Model)	Two triples *(h,r,t)* with r ∈ support, attack	0.604
Random Forest	Predictions from DeBERTa-v3 and DistMult	0.585
Gradient Boosting	Predictions from DeBERTa-v3 and DistMult	0.585
Decision Tree	Predictions from DeBERTa-v3 and DistMult	0.579
MultiLayer Perceptron	Predictions from DeBERTa-v3 and DistMult	0.579
K-Nearest Neighbors	Predictions from DeBERTa-v3 and DistMult	0.578

Table 7: Comparison of classifiers integrating DeBERTa-v3 and DistMult predictions for argument relation classification.

classify *support* as *no relation*. While DeBERTa-V3 handles *support* better than DistMult, the CNN Classifier outperforms both, achieving better balance. For *attack*, DeBERTa-V3 outperforms DistMult and the CNN. Overall, the CNN Classifier has the best balance across most classes, combining strengths and reducing misclassification.

This study shows KGEMs can enhance AM methods for argument relation prediction, particularly in political debates. While KGs have previously been applied in AM tasks, what is particularly novel in this work is their application to political debates using the *ElecDeb60to20* dataset. This dataset's diverse argumentation styles and topics present a challenging scenario for integrating KGEMs and LMs in this field. Other studies (Gemechu and Reed, 2019; Mestre et al., 2021; Ruiz-Dolz et al., 2021) have proposed different approaches on this task using subsets of the *ElecDeb60to20* dataset, like US2016 and US2020. Our work extensively evaluate our hybrid approach on the entire dataset, outperforming these competing approaches and standard baselines in classifying relations between arguments. These results make explicit the value of incorporating relational insights from knowledge graphs into AM tasks, particularly in domains as complex as political debates. By bridging the strengths of KGEMs and LMs, this study sets a new benchmark for argument relation prediction in highly challenging datasets.

6 Conclusion

This paper introduces a novel hybrid framework for predicting relations between argument components in argumentation graphs, combining structural insights from KGEMs with contextual understanding from fine-tuned LMs. We showed that KGEMs, despite their traditional use in KG's tasks, achieve competitive performance in argument relation prediction. Our experiments with DistMult demonstrate that structural knowledge alone captures meaningful relational patterns, achieving a Macro F1-Score of 0.60 on the challenging standard *ElecDeb60to20* benchmark for AM.

Integrating KGEMs with LMs significantly enhances the prediction accuracy. Using classifiers like Random Forests and CNNs to combine predictions, our approach achieved SOTA performance. Notably, we improved the Macro F1-Score to 0.68 with RoBERTa and further to 0.73 with DeBERTa-V3, representing a significant gain over prior SOTA methods (Goffredo et al., 2023a).

Our ensemble method integrates multiple models, highlighting the value of combining structural and contextual knowledge to improve AM tasks in complex domains like political debates.

Limitations

Our approach has been tested on the *ElecDeb60to20* dataset, which consists of U.S. presidential debates only. While this dataset is well-suited for our current study, it does not guarantee that the model will perform equally well on other types of debates, argumentative genres, or in different domains or languages. However, it is worth noticing that *ElecDeb60to20*, and more generally the political debates scenario, represent one of the most challenging argumentation data to test AM models against. The model's effectiveness may also be compromised by varying strategic communication styles across different countries or cultural contexts. We recognize the need for additional experiments across diverse datasets to assess and potentially improve the model's adaptability.

Lastly, while our method separates the training of KGEMs and LLMs, it does not fully leverage the potential benefits of integrated approaches. In future work, we plan to explore hybrid training approaches, such as KEPLER (Wang et al., 2021b), that concurrently optimize KGEMs and language modelling objectives, with the aim to further strengthen the alignment between argument structure and content.

Acknowledgments

We thank Pierre Monnin for his insights regarding this work. This work has been supported by the French government, through the 3IA Cote d'Azur Investments in the project managed by the National Research Agency (ANR) with the reference number ANR-23-IACL-0001.

References

Mehdi Ali, Max Berrendorf, Charles Tapley Hoyt, Laurent Vermue, Sahand Sharifzadeh, Volker Tresp, and Jens Lehmann. 2021. Pykeen 1.0: A python library for training and evaluating knowledge graph embeddings. *J. Mach. Learn. Res.*, 22:82:1–82:6.

Sakshi Arora, Ajay Rana, and Archana Singh. 2023. Argument mining: A categorical review. In *Modern Electronics Devices and Communication Systems: Select Proceedings of MEDCOM 2021*, pages 353–367. Springer.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2787–2795.

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: a data-driven analysis. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 5427–5433. ijcai.org.

Zongsheng Cao, Qianqian Xu, Zhiyong Yang, and Qingming Huang. 2022. ER: equivariance regularizer for knowledge graph completion. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 5512–5520. AAAI Press.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2d knowledge graph embeddings. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 1811–1818. AAAI Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Martin Drance, Fleur Mougin, Akka Zemmari, and Gayo Diallo. 2023. Pre-trained embeddings for enhancing multi-hop reasoning. In *International Joint Conference on Artificial Intelligence 2023 Workshop on Knowledge-Based Compositional Generalization*.

Stefano Faralli, Andrea Lenzi, and Paola Velardi. 2023. A benchmark study on knowledge graphs enrichment and pruning methods in the presence of noisy relationships. *J. Artif. Intell. Res.*, 78:37–68.

Debela Gemechu and Chris Reed. 2019. Decompositional argument mining: A general purpose approach

for argument graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 516–526, Florence, Italy. Association for Computational Linguistics.

Debela Gemechu, Ramon Ruiz-Dolz, and Chris Reed. 2024. ARIES: A general benchmark for argument relation identification. In *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)*, pages 1–14, Bangkok, Thailand. Association for Computational Linguistics.

Pierpaolo Goffredo, Elena Cabrio, Serena Villata, Shohreh Haddadan, and Jhonatan Torres Sanchez. 2023a. Disputool 2.0: A modular architecture for multi-layer argumentative analysis of political debates. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 16431–16433. AAAI Press.

Pierpaolo Goffredo, Mariana Espinoza, Serena Villata, and Elena Cabrio. 2023b. Argument-based detection and classification of fallacies in political debates. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 11101–11112. Association for Computational Linguistics.

Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2022. Fallacious argument classification in political debates. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4143–4149. ijcai.org.

Shohreh Haddadan, Elena Cabrio, and Serena Villata. 2019a. Disputool - A tool for the argumentative analysis of political debates. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 6524–6526. ijcai.org.

Shohreh Haddadan, Elena Cabrio, and Serena Villata. 2019b. Yes, we can! mining arguments in 50 years of US presidential campaign debates. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4684–4690, Florence, Italy. Association for Computational Linguistics.

Shohreh Haddadan, Elena Cabrio, and Serena Villata. 2019c. Yes, we can! mining arguments in 50 years of US presidential campaign debates. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4684–4690. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543.

Khalid Al Khatib, Yufang Hou, Henning Wachsmuth, Charles Jochim, Francesca Bonin, and Benno Stein. 2020. End-to-end argumentation knowledge graph construction. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7367–7374. AAAI Press.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Weichen Li, Patrick Abels, Zahra Ahmadi, Sophie Burkhardt, Benjamin Schiller, Iryna Gurevych, and Stefan Kramer. 2021. Topic-guided knowledge graph construction for argument mining. In *2021 IEEE International Conference on Big Knowledge, ICBK 2021, Auckland, New Zealand, December 7-8, 2021*, pages 315–322. IEEE.

Marco Lippi and Paolo Torroni. 2016. Argument mining from speech: Detecting claims in political debates. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 2979–2985. AAAI 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. *CoRR*, abs/1907.11692.

Eleonora Mancini, Federico Ruggeri, Andrea Galassi, and Paolo Torroni. 2022. Multimodal argument mining: A case study in political debates. In *Proceedings of the 9th Workshop on Argument Mining*, pages 158–170, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Tobias Mayer, Santiago Marro, Elena Cabrio, and Serena Villata. 2021. Enhancing evidence-based medicine with natural language argumentative analysis of clinical trials. *Artif. Intell. Medicine*, 118:102098.

Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2018. Never retreat, never retract: Argumentation analysis for political speeches. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4889–4896. AAAI Press.

Rafael Mestre, Razvan Milicin, Stuart E. Middleton, Matt Ryan, Jiatong Zhu, and Timothy J. Norman.

2021. M-arg: Multimodal argument mining dataset for political debates with audio and transcripts. In *Proceedings of the 8th Workshop on Argument Mining*, pages 78–88, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gaku Morio, Hiroaki Ozaki, Terufumi Morishita, and Kohsuke Yanai. 2022. End-to-end argument mining with cross-corpora multi-task learning. *Trans. Assoc. Comput. Linguistics*, 10:639–658.

Umer Mushtaq and Jérémie Cabessa. 2023. Argument mining with modular BERT and transfer learning. In *International Joint Conference on Neural Networks, IJCNN 2023, Gold Coast, Australia, June 18-23, 2023*, pages 1–8. IEEE.

Vlad Niculae, Joonsuk Park, and Claire Cardie. 2017. Argument mining with structured svms and rnns. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 985–995. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Prakash Poudyal, Jaromir Savelka, Aagje Ieven, Marie Francine Moens, Teresa Goncalves, and Paulo Quaresma. 2020. ECHR: Legal corpus for argument mining. In *Proceedings of the 7th Workshop on Argument Mining*, pages 67–75, Online. Association for Computational Linguistics.

David Powers. 2011. Evaluation: From precision, recall and f-measure to roc, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1):37–63.

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 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

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

Ameer Saadat-Yazdi, Jeff Z. Pan, and Nadin Kökciyan. 2023. Uncovering implicit inferences for improved relational argument mining. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 2476–2487. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Comput. Linguistics*, 43(3):619–659.

Jacky Visser, Barbara Konat, Rory Duthie, Marcin Koszowy, Katarzyna Budzynska, and Chris Reed. 2020a. Argumentation in the 2016 US presidential elections: annotated corpora of television debates and social media reaction. *Lang. Resour. Evaluation*, 54(1):123–154.

Jacky Visser, Barbara Konat, Rory Duthie, Marcin Koszowy, Katarzyna Budzynska, and Chris Reed. 2020b. Argumentation in the 2016 us presidential elections: annotated corpora of television debates and social media reaction. *Language Resources and Evaluation*, 54(1):123–154.

Meihong Wang, Linling Qiu, and Xiaoli Wang. 2021a. A survey on knowledge graph embeddings for link prediction. *Symmetry*, 13(3):485.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021b. KEPLER: A unified model for knowledge embedding and pre-trained language representation. *Trans. Assoc. Comput. Linguistics*, 9:176–194.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. arxiv. *arXiv preprint arXiv:1910.03771*.

Qi Yan, Jiaxin Fan, Mohan Li, Guanqun Qu, and Yang Xiao. 2022. A survey on knowledge graph embedding. In *7th IEEE International Conference on Data Science in Cyberspace, DSC 2022, Guilin, China, July 11-13, 2022*, pages 576–583. IEEE.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Jian Yuan, Zhongyu Wei, Donghua Zhao, Qi Zhang, and Changjian Jiang. 2021. Leveraging argumentation knowledge graph for interactive argument pair identification. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 2310–2319. Association for Computational Linguistics.

A LMs integration with KGEMs

To validate our approach, we conducted a comparative evaluation of several LMs to determine the most compatible with our architecture. Specifically, we evaluated DeBERTa-V3 (He et al., 2021), BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2019) and XLM-RoBERTa (Conneau et al., 2020). Results are reported in Table 8.

Model	Method	Macro F1 Score
DeBERTa	**seq-class**	**0.69**
BERT	sent-class	0.66
XLM-RoBERTa	seq-class	0.63
DistilBERT	seq-class	0.58

Table 8: Macro F1-Score of several LMs for the AM Relation Prediction Task

We integrated DeBERTa-V3 into our architecture due to its superior performance compared to other models.

B Hyperparameters

This section details the optimal hyperparameters identified for the models employed in this study. These configurations were determined through extensive experimentation and validation to achieve the best performance for each model.

B.1 RoBERTa

Following the methodology outlined by Goffredo et al. (2023a), RoBERTa was fine-tuned with a learning rate of $6e^{-5}$, a batch size of 8, and a maximum sentence length of 64 sub-word tokens per input example. The model was trained for 15 epochs.

B.2 DeBERTa-V3

The DeBERTa-V3 model achieved optimal performance with a learning rate of $4e^{-5}$, a batch size of 16, and a maximum sentence length of 255 sub-word tokens. It was fine-tuned over 3 epochs.

B.3 DistMult

The DistMult model's optimal configuration was obtained after 165 epochs. It used a learning rate of $1.35e^{-2}$, a batch size of 128, an embedding dimension of 160, and a margin ranking loss with a margin of 2.99.

B.4 Random Forest Classifier (RFC)

The RFC achieved its best performance using 50 estimators, the Gini criterion, a minimum of 2 samples required to split an internal node, and a minimum of 1 sample per leaf.

B.5 Convolutional Neural Network (CNN)

The CNN was evaluated using cross-validation with 30 epochs for each fold, a learning rate of $1e^{-3}$, an embedding dimension of 100 for the textual features, and a batch size of 32.