

A Dataset Independent Set of Baselines for Relation Prediction in Argument Mining

Oana Cocarascu¹, Elena Cabrio², Serena Villata², Francesca Toni¹

¹Imperial College London, UK

² Université Côte d’Azur, CNRS, Inria, I3S, France

{oana.cocarascu11, f.toni}@imperial.ac.uk, {elena.cabrio, serena.villata}@unice.fr

Abstract

Argument Mining is the research area which aims at extracting argument components and predicting argumentative relations (i.e., *support* and *attack*) from text. In particular, numerous approaches have been proposed in the literature to predict the relations holding between the arguments, and application-specific annotated resources were built for this purpose. Despite the fact that these resources have been created to experiment on the same task, the definition of a single relation prediction method to be successfully applied to a significant portion of these datasets is an open research problem in Argument Mining. This means that none of the methods proposed in the literature can be easily ported from one resource to another. In this paper, we address this problem by proposing a set of dataset independent strong neural baselines which obtain homogeneous results on all the datasets proposed in the literature for the argumentative relation prediction task. Thus, our baselines can be employed by the Argument Mining community to compare more effectively how well a method performs on the argumentative relation prediction task.

Keywords: Argument Mining; Discourse Annotation, Representation and Processing; Statistical Machine Learning Methods

1. Introduction

Argument(ation) Mining (AM) is “the general task of analyzing discourse on the pragmatics level and applying a certain argumentation theory to model and automatically analyze the data at hand” (Habernal and Gurevych, 2017). Two tasks are crucial (Peldszus and Stede, 2013; Lippi and Torroni, 2016; Cabrio and Villata, 2018): 1) argument component detection within the input natural language text aiming at the identification of arguments (claim, premises, and their textual boundaries); and 2) relation prediction aiming at the prediction of the relations between the argumentative components identified in the first stage (support, attack).

Despite the high volume of approaches tackling the relation prediction task with satisfying results (see (Cabrio and Villata, 2018) for the complete list), a problem arises: these solutions heavily rely on the peculiar features of the dataset taken into account for the experimental setting and are hardly portable from one application domain to another. On the one side, this issue can be explained by the huge number of heterogeneous application domains where argumentative text may be analysed (e.g., online reviews, blogs, political debates, legal cases). On the other side, it represents a drawback for the comparison of the different approaches proposed in the literature, which are often presented as solutions addressing the relation prediction task from a dataset independent point of view. A side drawback for the AM community is therefore a lack of big annotated resources for this task, as most of them cannot be successfully reused. In this paper, we tackle this issue by proposing a set of strong cross-dataset baselines based on different neural architectures. Our baselines are shown to perform homogeneously over all the datasets proposed in the literature for the relation prediction task in AM, differently from what is achieved by the single methods proposed in the literature. The contribution of our proposal is to bestow the

AM community with a set of strong cross-dataset baselines to compare with in order to demonstrate how well a relation prediction method for AM performs.

The majority of the datasets containing argumentative relations target only two types of relations: attack and support. We define neural models to address the binary classification problem, analysing, to the best of our knowledge, all available datasets for this task ranging from persuasive essays to user-generated content, to political speeches. Given two arguments, we are interested in determining the relation between the first, called *child* argument, and the second, called *parent* argument, by means of a neural network. For example, the child argument *People know video game violence is fake* attacks the parent argument *Youth playing violent games exhibit more aggression*. Each of the two arguments is represented using embeddings as well as other features.

Current papers that target AM propose different neural networks for different datasets. In this paper, we propose several neural network architectures and perform a systematic evaluation of these architectures on different datasets for the relation prediction in argument mining. We provide a broad comparison of different deep learning methods for a large number of datasets for the relation prediction in AM, an important and still widely open problem. Concretely, we propose four neural network architectures for the classification task, two concerned with the way child and parent are passed through the network (*concat* model and *mix* model), an autoencoder, and an attention-based model.

In the remainder of the paper, Section 2. presents the datasets used in the experiments, along with their main linguistic features. Section 3. describes the features, and the deep learning models. We report the performance of the proposed models in Section 4. Conclusions for the paper are in Section 5.

Dataset	ID	# attacks	# supports
Essays	essay	497	4841
Microtexts	micro	108	263
Nixon-Kennedy	nk	378	353
Debatepedia	db	141	179
IBM	ibm	1069	1325
ComArg	com	296	462
Web-content	web	1301	1329
CDCP	cdcp	0	1220
UKP	ukp	5935	4759
AIFdb	aif	9854	7543

Table 1: Summary of datasets.

2. Relation-based AM datasets

In this section, we describe the datasets that we used to compute our baselines¹. Datasets statistics can be found in Table 1². We focused on these datasets as they were specially created for the relation prediction in AM or they can be easily transformed to be used for this task.

- *Persuasive essays* (Stab and Gurevych, 2017): a corpus of 402 persuasive essays annotated with discourse-level argumentation structures. The major claim represents the author’s standpoint on the topic, which is supported or attacked by claims which in turn can be supported or attacked by premises. An example of (a part of) an essay is below:

Ever since researchers at the Roslin Institute in Edinburgh cloned an adult sheep, there has been an ongoing debate about whether cloning technology is morally and ethically right or not. Some people argue for and others against and there is still no agreement whether cloning technology should be permitted. However, as far as I’m concerned, [cloning is an important technology for humankind]_{MajorClaim1} since [it would be very useful for developing novel cures]_{Claim1}. First, [cloning will be beneficial for many people who are in need of organ transplants]_{Claim2}. [Cloned organs will match perfectly to the blood group and tissue of patients]_{Premise1} since [they can be raised from cloned stem cells of the patient]_{Premise2}.

In this example, both Claim1 and Claim2 support the Major Claim, Premise1 supports Claim2 and Premise2 supports Premise1.

- *Microtexts* (Peldszus and Stede, 2015): a corpus of 112 microtexts covering controversial issues. We focus on normal supports and rebut attacks only. The dataset has in addition examples and rebut attacks but we discard the former due to them being rarely used and the latter because we are not interested in attacks to inferences. An example of a microtext can be seen

¹We do not consider the two legal datasets built for relation prediction by (Mochales and Moens, 2011) and (Teruel et al., 2018) because the former is not available and the latter has a low inter-annotator agreement.

²For more details about the single datasets, we refer the reader to the related publication.

in Figure 1. Here, the second segment rebuts the first segment and the third segment undercuts the link between the second segment and the first segment. Segments four and five jointly support the main claim.

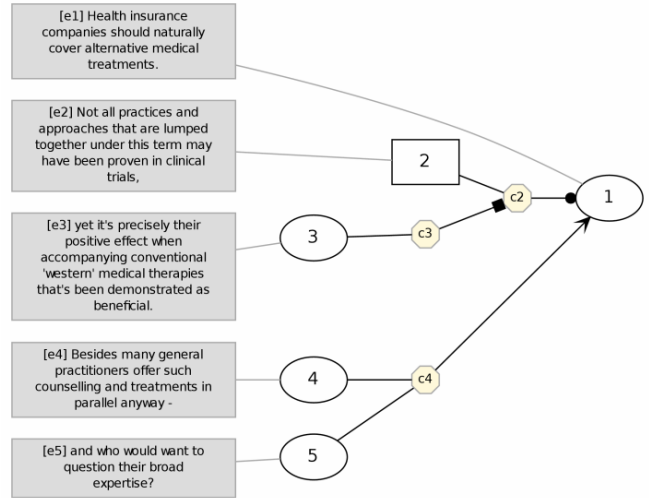


Figure 1: An example of microtext and the associated argumentation graph.

- *Nixon-Kennedy debate* (Menini et al., 2018): a corpus from the Nixon-Kennedy presidential campaign covering five topics: Cuba, disarmament, healthcare, minimum wage and unemployment. Below are two examples from the dataset:

*We could have tried to inject ourselves into the Congo without honoring our commitments to the United Nations charter, just as Khrushchev seems to be trying to do. We could have turned Cuba into a second Hungary. But we can be eternally grateful that we have a man in the White House who did none of these things **supports** I don’t take the views that the only alternative to a dictator is a Communist dictator. If the United States had just had its influence, and at that time the United States was extremely powerful in Cuba, it seems to me we could have persuaded Mr. Batista to hold free elections at the time he was permitted to go and permit the Cuban people to make their choice instead of letting Castro seize power through revolution. I think we are going to have a good deal of trouble in the future with Castro through all of Latin America.*

*They are afraid of diplomatic policies that teeter on the brink of war. They are dismayed that our negotiators have no solid plans for disarmament. And they are discouraged by a philosophy that puts its faith in swapping threats and insults with the Russians **attacks** It’s a typically specious and frivolous maneuver. We have made a good-faith effort to advance the - advance toward disarmament - and make some progress by having a meeting of the Disarmament Commission. Now, when they make a proposal like this, it’s a cynical attempt to prevent progress, that’s what it is it shows that they don’t really want disarmament.*

- *Debatepedia* (Cabrio and Villata, 2014): a corpus from the two debate platforms Debatepedia³ and ProCon⁴. Below are two examples from the dataset:

*Research studies have yielded the conclusion that the effect of violent media consumption on aggressive behavior is in the same ballpark statistically as the effect of smoking on lung cancer; the effect of lead exposure on children’s intellectual development and the effect of asbestos on laryngeal cancer. **supports** Violent video games are real danger to young minds.*

*People know video game violence is fake. **attacks** Youth playing violent games exhibit more aggression.*

- *IBM* (Bar-Haim et al., 2017): a dataset from topics randomly selected from the debate motions database at the International Debate Education Association (IDEA)⁵. Below are two examples from the dataset:

*Children with many siblings receive fewer resources. **supports** This house supports the one-child policy of the republic of China.*

*Virtually all developed countries today successfully promoted their national industries through protectionism **attacks** This house would unleash the free market.*

- *ComArg* (Boltužić and Šnajder, 2014): a corpus of online user comments from ProCon and IDEA. We combine the two types of attacks (explicit and vague/implicit attacks) and the two types of supports (explicit and vague/implicit arguments). Below are two examples from the dataset:

*Religion should stay out of the public square, except when people exercise their right to the freedom of speech an expression. Having Under God in the pledge forces all people to pledge allegiance to a higher power they may not believe in. The separation of Church and State should disallow such favoritism. Can anyone fathom the reaction of believers if it said: One Nation, created by a big bang and inhabited by evolved creatures.... ? **supports** Removing under god would promote religious tolerance*

*Atheism doesn’t mean the absence of religion - it means the absence of a god in one’s belief system. Certain forms of Buddhism, for example are atheistic. Therefore, requiring a statement of belief in a god is unconstitutionally preferring a majority religious belief over a minority one. The point of the Pledge is to state allegiance to the flag and country. If one believes in a god, there are many, many other forums in which to express that belief without imposing it on others. **attacks** America is based on democracy and the pledge should reflect the belief of the American majority.*

- *Web-content dataset* (Carstens and Toni, 2015): a dataset of arguments adapted from the Argument Corpus (Walker et al., 2012), plus arguments from news

articles, movies, ethics and politics. Below are two examples from the dataset:

*i agree did not like this either in fact i stopped watching once waltz was killed because i just didnt care anymore **supports** after all the attention and awards etc and an imdb rating of i was so shocked to finally see this film and have it be so bad*

- *samsung note it has a bigger screen and a somewhat faster processor **attacks** htc one it is currently the best one in the market good quality superb specs*

- *CDCP* (Park and Cardie, 2018): a dataset consisting of support arguments only from user comments regarding Consumer Debt Collection Practices from an eRulemaking website⁶. Below are two examples:

*sundays really are when most people are spending whatever little time they have left before the work-week with friends and family **supports** i do not conduct business on sundays*

- *a robo-call that tells you that you have a message or an account update, and the only way to get it is to call a special number with an extension, but when you call, it is just the same message asking where your payment is, is a waste of the consumer’s time and the consumer’s cellular resources (two phone calls, one received, one sent **supports** i support these restrictions on robo-calling and any calls during the work hours*

- *UKP* (Stab et al., 2018): a dataset of arguments from online comments on 8 controversial issues: abortion, cloning, death penalty, gun control, minimum wage, nuclear energy, school uniforms, marijuana legalization. In this dataset, one of the arguments is represented by the topic. Below are two examples:

*Dr. Strouse has seen both the benefits and risks of cannabis use and is well-versed in the emerging scientific evidence regarding the effectiveness of cannabinoids in a variety of medical conditions and pain states, as well as epidemiologic evidence of legalized marijuana’s connection to a reduction in prescription drug use and opioid-related deaths **supports** marijuana legalization*

*Would you want to live in a neighborhood filled with people who regularly smoke marijuana **attacks** marijuana legalization*

For our experiments, we modify the parent text from *topic* to a default seen as the natural language template *topic is good*. Hence from the previous example, we would have an argument for and an argument against “marijuana legalization is good”.

- *AIFdb* (Bex et al., 2013; Chesñevar et al., 2006; Iyad and Reed, 2009; Reed et al., 2008): a corpus of argument maps which follows the structure defined by AIF (Lawrence et al., 2012). We select the following datasets from AIFdb and keep the English texts only: AraucariaDB, DbyD Argument Study, Expert

³<http://idebate.org/deATABASE>

⁴<http://www.procon.org/>

⁵<http://idebate.org/>

⁶<http://regulationroom.org>

Opinion and Positive Consequences, Internet Argument Corpus, Mediation (here we compiled the following datasets: Dispute mediation, Dispute mediation: excerpts taken from publications, Mock mediation, Therapeutic, Bargaining, Meta-talk in mediation), Opposition (here we compiled the following datasets: Language Of Opposition Corpus 1, Android corpus, Ban corpus, Ipad corpus, Layoffs corpus, Twitter corpus). We map the original set of relations to 2 classes as follows: CA-nodes are mapped to attack and RA- and TA-nodes are mapped to support. Below are two examples from the dataset:

the water temperature is perfect **supports** *Burleigh Heads Beach is the best.*

We should implement Zoho, because it is cheaper than MS Office **attacks** *We should implement OpenOffice.*

In terms of results reported on the datasets we have conducted our experiments on, most works perform a cross-validation evaluation or, in the case of datasets consisting of several topics, the models proposed are trained on some of the topics and tested on the remaining topics.

For the *essay* dataset, an Integer Linear Programming model was used to achieve 0.947 F_1 for the support class and 0.413 F_1 for the attack class on the testing dataset using cross-validation to select the model (Stab and Gurevych, 2017). Using SVM, 0.946 F_1 for the support class and 0.456 F_1 for the attack class were obtained (Stab and Gurevych, 2017). Using a modification of the Integer Linear Programming model to accommodate the lack of some features used for the *essay* dataset but not present in the *micro* dataset, 0.855 F_1 was obtained for the support class and 0.628 F_1 for the attack class. On the *micro* dataset, an evidence graph model was used to achieve 0.71 F_1 using cross-validation (Peldszus and Stede, 2015). On the *nk* dataset, 0.77 F_1 for the attack class and 0.75 F_1 for the support class were obtained using SVM and cross-validation (Menini et al., 2018). SVM accuracy results on the testing dataset using coverage (i.e. number of claims identified over the number of total claims) were reported in (Bar-Haim et al., 2017) as follows: 0.849 accuracy for 10% coverage, 0.740 accuracy for 60% coverage, 0.632 accuracy for 100% coverage. Random Forests were evaluated on the *web* and *aif* datasets using cross-validation, achieving 0.717 F_1 and 0.831 F_1 , respectively (Carstens and Toni, 2017). Structured SVMs were evaluated in a cross-validation setting on the *cdep* and *ukp* datasets using various types of factor graphs, full and strict (Niculae et al., 2017). On the *cdep* dataset, F_1 was 0.493 on the full graph and 0.50 on the strict graph whereas on the *ukp* dataset, F_1 was 0.689 on the full graph and 0.671 on the strict graph. No results on the two-class datasets were reported for *db*, *com*, and *ukp* datasets. The results on *ukp* treat either supporting and attacking arguments as a single category or considering three types of relations: support, attack, neither. The latter type of reporting results on three classes is also given on the *com* dataset.

3. Neural baselines for relation prediction

In this section we describe the features used and the proposed neural models.

3.1. Features

We use four types of features: embeddings, textual entailment, sentiment features, and syntactic features, computed for *child* and *parent*, respectively. We refer to the last three types of features as *standard* features.

Word embeddings are distributed representations of texts in an n -dimensional space. We add a feature of entailment from child to parent representing the class (entailment, contradiction, or neutral) obtained using AllenNLP⁷, a textual entailment model based on a decomposable attention model (Parikh et al., 2016). The features related to sentiment are based on manipulation of SentiWordNet (Esuli and Sebastiani, 2006) and the sentiment of the entire text analysed using the VADER sentiment analyser (Hutto and Gilbert, 2014). Every WordNet synset (Miller, 1995) can be associated to three scores describing how objective, positive, and negative it is. For every word in the text (child and parent, respectively), we select its first synset and compute its positive score and its negative score. In summary, the features related to sentiment for a text t that consists of n words, $i=1..n$, are the following: (i) sentiment score ($\sum_{w_i} pos_score(w_i) - neg_score(w_i)$), (ii) number of positive/negative/neutral words in t (a word is neutral if $not(pos_score(w_i) > 0 \text{ and } neg_score(w_i) > 0)$), (iii) sentiment polarity class and score of t . Syntactic features consist of text statistics and word statistics with respect to the POS tag: number of words, nouns, verbs, first person singular, second person singular and plural, third person singular and plural, first person plural, modals, modifiers (number of adverbs plus the number of adjectives), and lexical diversity (number of unique words divided by the total number of words in text t).

3.2. Neural Architectures

We describe the four neural architectures we propose for determining the argumentative relation (attack or support) holding between two texts (see Figures 1-4). For all, the number of the hidden layers and their sizes are the ones that performed the best. We report only on configurations of the architectures as given in Section 3.2. as these were the best performing. However, we experimented with 1 and 2 hidden layers, and hidden layer sizes of 32, 64, 128, and 256, trying all possible combinations in order to obtain the best configurations, and limiting to 2 hidden layers due to the small size of the data. For our models, we use GRUs (Cho et al., 2014). Various works have compared LSTMs and GRUs but (Chung et al., 2014; Józefowicz et al., 2015) did not obtain conclusive results as to which type is better, suggesting that the design choice is dependant on the dataset and task. We focus on GRUs as they take less time to train and are more efficient as LSTMs have more parameters.

3.2.1. Concat model

In the concat model, each of the child and parent embeddings is passed through a GRU. We concatenate the standard features of the child and parent. The merged standard vector is then concatenated with the outputs of the GRUs. The resulting vector is passed through 2 dense layers (of 256 neurons and 64 neurons, with sigmoid as activation

⁷<https://allennlp.org/models>

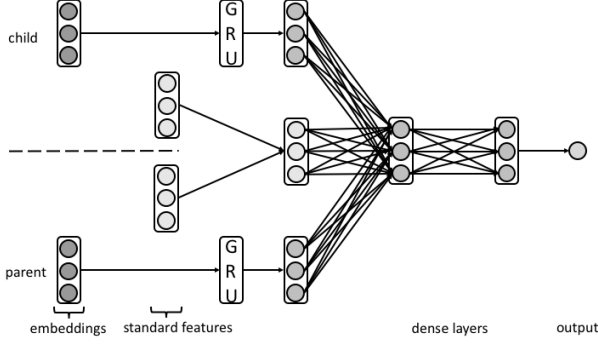


Figure 2: The concat architecture.

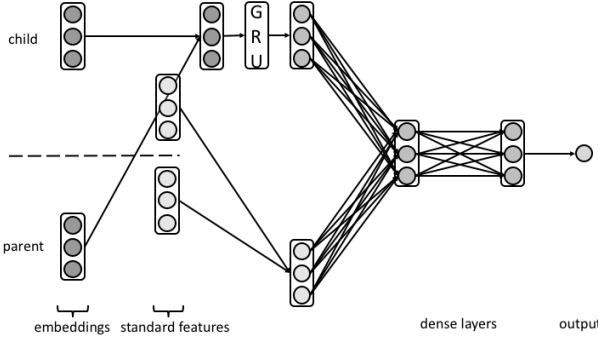


Figure 3: The mix architecture.

function), and then to softmax to determine the argumentative relation. The concat model can be seen in Figure 2.

3.2.2. Mix model (M)

In the mix model, we first concatenate the child and parent embeddings and then pass them through a GRU, differently from the concatenation model where we pass each embedding vector through a GRU first. We concatenate the standard features that we obtain for the child and for the parent, respectively. The merged standard vector is then concatenated with the output of the GRU. From this stage, the network resembles the concatenation model: the resulting vector is passed through 2 dense layers (of 256 neurons and 64 neurons, with sigmoid as activation function), to be then finally passed to softmax to determine the argumentative relation. The mix model can be seen in Figure 3.

3.2.3. Autoencoder model

Autoencoders (Hinton and Salakhutdinov, 2006; Erhan et al., 2010) are unsupervised learning neural models which take a set of features as input and aim, through training, to reconstruct the inputs. Autoencoders can be used as feature selection methods to determine which features are redundant (Wang et al., 2017; Han et al., 2017). We first concatenate the child and parent tensors, to obtain a vector of size X . We use an autoencoder with one hidden layer defined as: (i) an encoder function $f(X) = \sigma(XW^{(1)})$, and (ii) a decoder function $\sigma(f(X)W^{(2)})$, where $W^{(1)}, W^{(2)}$ are the weight parameters in the encoder and decoder, respectively. The size of the hidden layer is 128. We use sigmoid as activation function in the autoencoder and binary cross entropy as loss function. We concatenate the standard features of

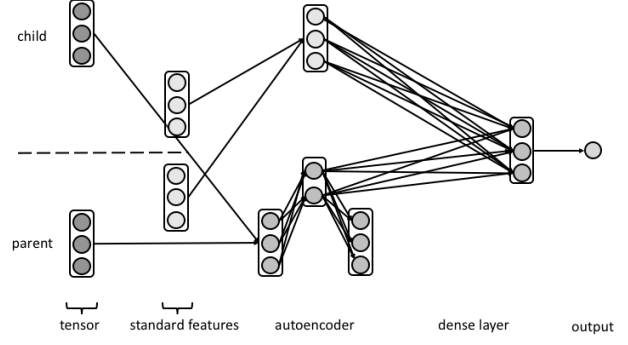


Figure 4: The autoencoder architecture.

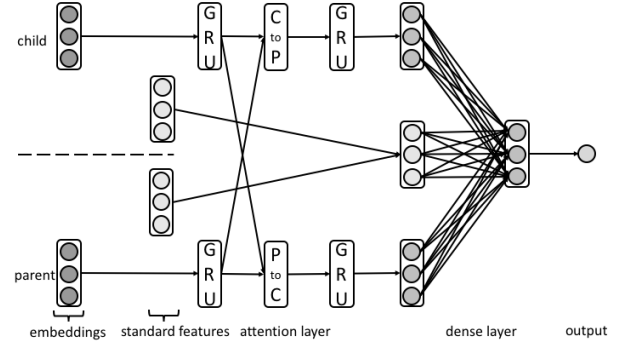


Figure 5: The attention architecture.

the child and of the parent. The merged standard vector is then concatenated with the hidden layer in the autoencoder (Figure 4) which represents the encoded dataset as dimensionally reduced features. The resulting vector is passed through a single dense layers (of 32 neurons, with sigmoid as activation function), that is then passed to softmax.

3.2.4. Attention model

Inspired by the demonstrated effectiveness of attention-based models (Yang et al., 2016; Vaswani et al., 2017), we combine the GRU-based model with attention mechanisms. Each of the child and parent embeddings is passed through a GRU. Let $C \in \mathbb{R}^{L_c \times d}$ be the output the GRU produces when reading L_c words of Child C , and let $P \in \mathbb{R}^{L_p \times d}$ be the output the GRU produces when reading L_p words of Parent P , where d is the output dimension. We compute attention in two directions: from Child C to Parent P and from P to C . We illustrate the attention in one direction only. Let s_{ij} be the similarity matrix between the i -th child word and the j -th parent word, α the attention weight, c'_i the attended child vector, and c''_i parent-aware representation of each child as follows:

$$\begin{aligned}
 s_{ij} &= W^{1 \times 2d} [c_i; p_j]^{2d \times 1} + b^{1 \times 1} & s &\in \mathbb{R}^{L_c \times L_p} \\
 \alpha_i &= \text{softmax}(s_{i:}) & \alpha &\in \mathbb{R}^{L_c \times L_p} \\
 c'_i &= \alpha_i \cdot p_i & c' &\in \mathbb{R}^{L_c \times d} \\
 c''_i &= [c_i; c'_i] & c'' &\in \mathbb{R}^{L_c \times 2d}
 \end{aligned}$$

where W is a trainable weight vector and $[\cdot]$ is vector concatenation across row. The weights vectors W for the two directions are different. We concatenate the standard features of the child and of the parent. The merged standard

vector is then concatenated with the outputs of the GRUs whose inputs are c'' and p'' . The resulting vector is passed through a single dense layers (128 neurons, activation function: sigmoid), that is then passed to softmax (Figure 5).

4. Experimental results

4.1. Non-neural baselines

We have used for training the larger datasets, *aif*, *essay*, *ibm* and *web*. We resampled the minority class from the *essay* dataset and used our models on the oversampled dataset. We did not use for training the *ukp* dataset as the parent is a topic instead of an argument. The models are then tested on the remaining datasets with the average being computed on testing datasets. We report the F_1 performance of the *attack* class (A) and the *support* class (S). Table 2 shows the results for the non-neural baselines. We used Random Forests (RF) (Breiman, 2001) with 15 trees in the forest and gini impurity criterion and SVM with linear kernel using LIBSVM (Chang and Lin, 2011), obtained as a result of performing a grid search as it is the most commonly used algorithm in the works that experiment on the datasets we considered (Bar-Haim et al., 2017; Boltužić and Šnajder, 2014; Carstens and Toni, 2017; Menini et al., 2018; Niculae et al., 2017). On top of the *standard* features used for our neural models, for the baselines we added the following features: TF-IDF, number of common nouns, verbs and adjectives between the two texts as in (Menini et al., 2018), a different sentiment score $\frac{nr_pos - nr_neg}{nr_pos + nr_neg + 1}$ as in (Bar-Haim et al., 2017), all features being normalized.

4.2. Neural baselines with non-contextualised word embeddings

Table 3 shows the best baselines for relation prediction in AM. For our models, we experimented with two types of embeddings: GloVe (Pennington et al., 2014) (300-dimensional) and FastText (FT) (Joulin et al., 2016; Mikolov et al., 2018) (300-dimensional). We used pre-trained word representations in all our models. We used 100 as the sequence size as we noticed that there are few instances with more than 100 words. We used a batch size of 32 and trained for 10 epochs (as a higher number of epochs led to overfitting). We report the results using embeddings and *syntactic* features and the results with *all* the features presented in Section 3.1. We also conducted a feature ablation experiment (with embeddings being always used) and observed that syntactic features contribute the most to performance, with the other types of features bringing small improvements when used together only with embeddings. In addition, we have run experiments using two datasets for training to test whether combining two datasets improves performance. During training, we used one of the large datasets (*aif*, *essay*, *ibm*, *web*) and one of the remaining datasets (represented as blanks in the Table⁸).

Amongst the proposed architectures, the attention model generally performs better. Using GloVe embeddings instead of FastText yields better results. The autoencoder does not give good results, which may be attributed to the fact that using the encoding of the most salient features is

not enough in predicting the argumentative relation and that analysing the entire sequence is better. Using only a single dataset for training, the models that perform the best are the *attention* model and the *mix* model, in both cases using *all* features and trained on the *essay* dataset. The best results are obtained when using another dataset along one of the larger datasets for training. This is because combining data from two domains we are able to learn better the types of argumentative relations. When using *syntactic* features, adding *micro*, *cdcp*, and *ukp* does not improve the results compared to using a single dataset for training. Indeed, *cdcp* has only one type of relation (i.e. support) resulting in an imbalanced dataset and in *ukp*, the parent argument is a topic, which does not improve the prediction task. When using *all* features, *micro*, *com*, *ukp*, and *nk* do not contribute to an increase in performance. The best results are obtained using the attention mechanism with GloVe embeddings trained on the *web* and *essay* datasets using *syntactic* features (0.5445 macro average F_1).

4.3. Neural baselines with contextualised word embeddings

Contextualised word embeddings such as the Bidirectional Encoder Representations from Transformers (BERT) embeddings (Devlin et al., 2018) analyse the entire sentence before assigning an embedding to each word. The main difference between GloVe, FastText and contextualised word embeddings is that GloVe does not take the word order into account during training, whereas BERT do. We employ BERT embeddings to test whether they bring any improvements to the classification task. While for GloVe/FastText vectors we do not need the original, trained model in order to use the embeddings, for the contextualised word embeddings we require the pre-trained language models that we can then fine tune using the datasets of the downstream task. We try different combinations for the neural network with BERT embeddings: using 3 or 4 BERT layers and using 1 dense layer (of 64 neurons) or 2 dense layers (of 128 and 32 neurons) before the final layer that determines the class. Table 4 shows the results with BERT embeddings instead of GloVe/FastText, following the same experiments described in Section 4.2.: feature ablation (*syntactic* vs *all* features) and using two datasets for training to test whether this can improve performance. The best results are obtained using 4 BERT layers and 2 dense layers (0.537 macro average F_1). However, the best BERT baseline does not outperform the best results obtained using the attention model and GloVe.

4.4. Discussion

Our baselines perform homogeneously over all existing datasets for relation prediction in AM while using generic features. As it may be noticed in the examples provided in Section 2., the datasets differ at granularity: some consist of pairs of sentences (e.g., IBM) whereas others include pair of multiple-sentence arguments (e.g., Nixon-Kennedy debate). Additionally, the argumentation relations can be domain-specific and the semantic nature of argumentative relations may vary between corpora (e.g., ComArg). Thus, in this paper we considered a simpler but still complex task of determining the relation of either support or attack be-

⁸For readability, blanks represent the training datasets.

			essay	micro	db	ibm	com	web	cdcp	ukp	nk	aif	Avg	Mcr Avg
non-neural baselines	RF	F_1 A		0.24	0.22	0.25	0.03	0.27	-	0.22	0.43	0.31	0.246	0.467
		F_1 S		0.80	0.71	0.67	0.75	0.63	0.94	0.60	0.53	0.56	0.688	
	RF	F_1 A	0.32	0.24	0.40		0.33	0.38	-	0.39	0.55	0.44	0.381	0.508
		F_1 S	0.57	0.74	0.64		0.67	0.59	0.85	0.53	0.59	0.54	0.636	
	RF	F_1 A	0.57	0.40	0.45	0.53	0.43		-	0.60	0.52	0.57	0.509	0.490
		F_1 S	0.44	0.47	0.52	0.41	0.57		0.51	0.45	0.50	0.38	0.472	
	RF	F_1 A	0.67	0.45	0.60	0.62	0.56	0.62	-	0.71	0.68		0.614	0.335
		F_1 S	0.01	0.02	0.17	0.01	0.04	0.24	0.01	0.00	0.00		0.056	
SVM	F_1 A		0.34	0.36	0.33	0.29	0.38	-	0.42	0.42	0.40	0.368	0.503	
	F_1 S		0.71	0.67	0.65	0.67	0.59	0.84	0.57	0.56	0.49	0.639		
SVM	F_1 A	0.52	0.42	0.37		0.43	0.50	-	0.59	0.33	0.49	0.456	0.473	
	F_1 S	0.48	0.51	0.50		0.45	0.47	0.61	0.43	0.53	0.42	0.489		
SVM	F_1 A	0.49	0.35	0.39	0.39	0.38		-	0.56	0.57	0.520	0.456	0.498	
	F_1 S	0.50	0.54	0.52	0.59	0.60		0.67	0.46	0.47	0.500	0.539		
SVM	F_1 A	0.61	0.40	0.60	0.46	0.57	0.61	-	0.64	0.68		0.571	0.431	
	F_1 S	0.35	0.50	0.22	0.57	0.04	0.24	0.40	0.30	0.00		0.291		

Table 2: Results on the datasets with attack (A) and support (S) relations. F_1 A stands for the F_1 measure of the attack relation and F_1 S stands for the F_1 measure of the support (S) relation. RF stands for Random Forests. The blanks represent the training dataset. The Average (Avg) and the Macro (Mcr) Avg do not include the results of the dataset used for training.

			essay	micro	db	ibm	com	web	cdcp	ukp	nk	aif	Avg	Mcr Avg
embeddings + syntactic	M FT	F_1 A	0.52	0.40	0.58	0.50	0.52		-	0.58	0.52		0.517	0.521
		F_1 S	0.47	0.50	0.61	0.54	0.54		0.60	0.42	0.53		0.526	
	C FT	F_1 A		0.36		0.45	0.47	0.52	-	0.65	0.46	0.52	0.490	0.52
		F_1 S		0.64		0.55	0.67	0.53	0.72	0.34	0.48	0.47	0.550	
	C G	F_1 A		0.35	0.43	0.48	0.31	0.45	-	0.58		0.43	0.433	0.526
F_1 S			0.71	0.68	0.58	0.70	0.54	0.77	0.47		0.50	0.619		
A G	F_1 A		0.37	0.58		0.53	0.53	-	0.61	0.59	0.55	0.537	0.526	
	F_1 S		0.61	0.60		0.42	0.50	0.72	0.43	0.38	0.47	0.516		
A G	F_1 A		0.36	0.48	0.43	0.39		-	0.52	0.45	0.51	0.449	0.544	
	F_1 S		0.75	0.66	0.62	0.68		0.79	0.52	0.56	0.54	0.640		
all features	M G	F1 0	0.49	0.41	0.51		0.50	0.52	-	0.57	0.49		0.499	0.517
		F1 1	0.49	0.63	0.60		0.52	0.45	0.69	0.37	0.54		0.536	
	C G	F1 0	0.42	0.33	0.52	0.35	0.48		-	0.49	0.43		0.431	0.515
		F1 1	0.54	0.68	0.63	0.66	0.53		0.75	0.46	0.55		0.600	
	A G	F1 0		0.30	0.42	0.40	0.38	0.43	-	0.53	0.35	0.48	0.411	0.515
		F1 1		0.73	0.61	0.60	0.68	0.55	0.82	0.49	0.56	0.53	0.619	
	M G	F1 0		0.37	0.43	0.43	0.40	0.46	-	0.71	-	0.46	0.466	0.532
		F1 1		0.71	0.64	0.61	0.70	0.55	0.78	0.11	0.78	0.51	0.599	
	C FT	F1 0		0.37		0.50	0.49	0.54	-	0.68	0.51	0.57	0.523	0.509
		F1 1		0.59		0.50	0.59	0.50	0.68	0.24	0.43	0.43	0.495	
A G	F1 0		0.40	0.51	0.52	0.45	0.54		0.60	0.56	0.59	0.521	0.512	
	F1 1		0.61	0.57	0.52	0.61	0.45		0.42	0.42	0.43	0.504		
A G	F1 0		0.36	0.54		0.50	0.51	-	0.59	0.59	0.55	0.520	0.535	
	F1 1		0.67	0.63		0.49	0.51	0.74	0.47	0.41	0.49	0.551		
A FT	F1 0		0.40	0.43	0.43	0.37		-	0.55	0.26	0.50	0.420	0.522	
	F1 1		0.72	0.64	0.58	0.72		0.77	0.47	0.59	0.51	0.625		
A G	F1 0		0.43	0.54	0.49	0.46		-	0.59	0.63	0.63	0.539	0.539	
	F1 1		0.68	0.55	0.57	0.56		0.65	0.46	0.38	0.47	0.540		

Table 3: Results on the datasets with attack (A) and support (S) relations. F_1 A stands for the F_1 measure of the attack relation and F_1 S stands for the F_1 measure of the support (S) relation. A stands for autoencoder model, C for concatenation model, M for mix model, G for GloVe embeddings, and FT for FastText embeddings. The blanks represent the training datasets. The Average (Avg) and the Macro (Mcr) Avg do not include the results of the dataset(s) used for training.

			essay	micro	db	ibm	com	web	cdep	ukp	nk	aif	Avg	Mcr Avg
BERT embeddings + syntactic	3B	F_1 A			0.53	0.48	0.52	0.49	-	0.56	0.46	0.43	0.496	0.522
	1D	F_1 S			0.63	0.56	0.58	0.50	0.70	0.48	0.51	0.42	0.548	
	4B	F_1 A			0.55	0.47	0.53	0.50	-	0.56	0.48	0.45	0.506	0.526
	2D	F_1 S			0.61	0.57	0.59	0.49	0.69	0.47	0.48	0.46	0.545	
	4B	F_1 A		0.36	0.48	0.40	0.45	0.42	-	0.53		0.37	0.430	0.525
	1D	F_1 S		0.69	0.67	0.61	0.62	0.57	0.79	0.50		0.50	0.619	
	3B	F_1 A		0.39	0.57	0.53	0.46	0.53	-	0.61	0.57		0.523	0.520
	1D	F_1 S		0.59	0.57	0.44	0.54	0.49	0.61	0.36	0.53		0.516	
	4B	F_1 A		0.37	0.54	0.52	0.43	0.51	-	0.58	0.56		0.501	0.521
	2D	F_1 S		0.61	0.58	0.45	0.57	0.52	0.65	0.40	0.55		0.541	
	3B	F_1 A		0.30	0.53		0.51	0.39	-	0.44	0.44	0.47	0.440	0.525
	2D	F_1 S		0.72	0.64		0.61	0.57	0.80	0.52	0.54	0.47	0.609	
	4B	F_1 A		0.33	0.49		0.49	0.40	-	0.56	0.47	0.44	0.454	0.531
	1D	F_1 S		0.68	0.66		0.61	0.56	0.78	0.46	0.56	0.55	0.608	
	4B	F_1 A		0.29	0.49		0.50	0.33	-	0.43	0.40	0.36	0.400	0.522
	2D	F_1 S		0.72	0.68		0.64	0.59	0.83	0.53	0.57	0.59	0.644	
3B	F_1 A	0.49	0.35	0.47		0.53	0.46		0.63	0.54	0.62	0.511	0.521	
2D	F_1 S	0.53	0.63	0.55		0.64	0.50		0.35	0.53	0.51	0.530		
4B	F_1 A	0.50	0.36	0.46		0.50		-	0.52	0.47	0.50	0.473	0.537	
2D	F_1 S	0.61	0.62	0.59		0.61		0.74	0.52	0.50	0.61	0.600		
4B	F_1 A		0.39	0.54	0.47	0.52		-	0.58	0.50	0.58	0.511	0.520	
1D	F_1 S		0.61	0.55	0.52	0.59		0.69	0.46	0.48	0.32	0.528		
all features	3B	F_1 A		0.33	0.42	0.41	0.52	0.42	-	0.53	0.45	0.35	0.429	0.524
	1D	F_1 S		0.67	0.66	0.63	0.63	0.58	0.80	0.52	0.54	0.53	0.618	
	3B	F_1 A			0.53	0.49	0.51	0.49	-	0.58	0.48	0.45	0.504	0.522
	1D	F_1 S			0.62	0.56	0.61	0.50	0.69	0.46	0.48	0.40	0.540	
	4B	F_1 A			0.53	0.50	0.54	0.51	-	0.59	0.51	0.49	0.524	0.529
	1D	F_1 S			0.59	0.56	0.55	0.47	0.67	0.45	0.48	0.49	0.533	
	3B	F_1 A	0.48	0.34	0.48		0.45		-	0.45	0.50	0.54	0.463	0.532
	2D	F_1 S	0.57	0.65	0.60		0.64		0.73	0.55	0.52	0.54	0.600	

Table 4: Results on the datasets with attack (A) and support (S) relations. F_1 A stands for the F_1 of the attack relation and F_1 S stands for the F_1 of the support (S) relation. XB stands for the number of BERT layers used (i.e. X) and YB stands for the number of dense layers (i.e. Y) used before the final layer that predicts the class. The blanks represent the training datasets. The Average (Avg) and the Macro (Mcr) Avg do not include the results of the dataset(s) used for training.

tween two texts. Embeddings represent the main difference in the features used for the machine learning models we experimented with. Whilst word embeddings are often used as the first data processing layer in a deep learning model, we employed TF-IDF features for the standard machine learning models that we considered as baselines. Other works that address the task of relation prediction make use of features specific to the single dataset of interest, making it difficult to test those models on the other datasets. For instance, for the *essay* dataset, Stab and Gurevych (2017) use structural features such as number of preceding and following tokens in the covering sentence, number of components in paragraph, number of preceding and following components in paragraph, relative position of the argument component in paragraph. For the other datasets, (Stab et al., 2018) use topic similarity features (as the *parent* argument is a topic), (Menini et al., 2018) use the position of the topic and similarity with other related/unrelated pair from the dataset, keyword embeddings of topics from the dataset. We have used only general purpose features that are meaningful for all datasets addressing the relational AM task. Surprisingly, BERT embeddings that have achieved state-

of-the-art in several tasks (Devlin et al., 2018) do not bring any improvements compared to non-contextualised word embeddings for the relation prediction task in AM.

5. Conclusion

Several resources have been built in the latest years for the task of argumentative relation prediction, covering different topics like political speeches, Wikipedia articles, persuasive essays. Given the heterogeneity of these kinds of text, it is hard to compare cross-dataset the different approaches proposed in the literature to address the argumentative relation prediction task. For this reason, in this paper, we addressed the issue of AM models that are hardly portable from one application dataset to another due to the features used. We provided a broad comparison of different deep learning methods using both non-contextualised and contextualised word embeddings for a large set of datasets for the argumentative relation prediction, an important and still widely open problem. We proposed a set of strong dataset independent baselines based on several neural architectures and have shown that our models perform homogeneously over all existing datasets for relation prediction in AM.

6. Bibliographical References

- Bar-Haim, R., Bhattacharya, I., Dinuzzo, F., Saha, A., and Slonim, N. (2017). Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261.
- Bex, F., Modgil, S., Prakken, H., and Reed, C. (2013). On logical specifications of the argument interchange format. *Journal of Logic and Computation*, 23(5):951–989.
- Boltužić, F. and Šnajder, J. (2014). Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.
- Cabrio, E. and Villata, S. (2014). Node: A benchmark of natural language arguments. In *Computational Models of Argument - Proceedings of COMMA 2014, Atholl Palace Hotel, Scottish Highlands, UK, September 9-12, 2014*, pages 449–450.
- Cabrio, E. and Villata, S. (2018). Five years of argument mining: a data-driven analysis. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI*, pages 5427–5433.
- Carstens, L. and Toni, F. (2015). Towards relation based argumentation mining. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 29–34.
- Carstens, L. and Toni, F. (2017). Using argumentation to improve classification in natural language problems. *ACM Transactions on Internet Technology*, 17(3):30:1–30:23.
- Chang, C. and Lin, C. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology TIST*, 2(3):27:1–27:27.
- Chesñevar, C. I., McGinnis, J., Modgil, S., Rahwan, I., Reed, C., Simari, G. R., South, M., Vreeswijk, G., and Willmott, S. (2006). Towards an argument interchange format. *Knowledge Engineering Review*, 21(4):293–316.
- Cho, K., van Merriënboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN Encoder–Decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.
- Chung, J., Gülçehre, Ç., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555.
- Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.
- Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P., and Bengio, S. (2010). Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 11:625–660.
- Esuli, A. and Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC’06)*, pages 417–422.
- Habernal, I. and Gurevych, I. (2017). Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.
- Han, K., Li, C., and Shi, X. (2017). Autoencoder feature selector. *CoRR*, abs/1710.08310.
- Hinton, G. and Salakhutdinov, R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504 – 507.
- Hutto, C. J. and Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM*.
- Iyad, R. and Reed, C. (2009). The argument interchange format. In *Argumentation in Artificial Intelligence*, pages 383–402.
- Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., and Mikolov, T. (2016). Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.
- Józefowicz, R., Zaremba, W., and Sutskever, I. (2015). An empirical exploration of recurrent network architectures. In *Proceedings of the 32nd International Conference on Machine Learning, ICML*, pages 2342–2350.
- Lawrence, J., Bex, F., Reed, C., and Snaith, M. (2012). Aifdb: Infrastructure for the argument web. In *Computational Models of Argument - Proceedings of COMMA*, volume 245, pages 515–516.
- Lippi, M. and Torroni, P. (2016). Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology*, 16(2):10.
- Menini, S., Cabrio, E., Tonelli, S., and Villata, S. (2018). Never retreat, never retract: Argumentation analysis for political speeches. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 4889–4896.
- Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. (2018). Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.
- Miller, G. A. (1995). Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41.
- Mochales, R. and Moens, M.-F. (2011). Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.
- Niculae, V., Park, J., and Cardie, C. (2017). Argument mining with structured SVMs and RNNs. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 985–995.
- Parikh, A. P., Täckström, O., Das, D., and Uszkoreit, J. (2016). A decomposable attention model for natural language inference. In *EMNLP*.
- Park, J. and Cardie, C. (2018). A corpus of eRulemaking user comments for measuring evaluability of arguments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC*.
- Peldszus, A. and Stede, M. (2013). From argument di-

- agrams to argumentation mining in texts: A survey. *IJCINI*, 7(1):1–31.
- Peldszus, A. and Stede, M. (2015). Joint prediction in mst-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 938–948.
- Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.
- Reed, C., Wells, S., Devereux, J., and Rowe, G. (2008). AIF+: dialogue in the argument interchange format. In *Computational Models of Argument: Proceedings of COMMA*, pages 311–323.
- Stab, C. and Gurevych, I. (2017). Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.
- Stab, C., Miller, T., Schiller, B., Rai, P., and Gurevych, I. (2018). Cross-topic argument mining from heterogeneous sources. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Teruel, M., Cardellino, C., Cardellino, F., Alemany, L. A., and Villata, S. (2018). Increasing argument annotation reproducibility by using inter-annotator agreement to improve guidelines. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC*.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*, pages 6000–6010.
- Walker, M. A., Tree, J. E. F., Anand, P., Abbott, R., and King, J. (2012). A corpus for research on deliberation and debate. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC*, pages 812–817.
- Wang, S., Ding, Z., and Fu, Y. (2017). Feature selection guided auto-encoder. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 2725–2731.
- Yang, Z., Yang, D., Dyer, C., He, X., Smola, A. J., and Hovy, E. H. (2016). Hierarchical attention networks for document classification. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.