Please quote as: Wambsganss, T., and Nikolaos Molyndris,, and Söllner, M. (2020) Unlocking Transfer Learning in Argumentation Mining: A Domain-Independent Modelling Approach. In 15th International Conference on Wirtschaftsinformatik (WI), pp 341-356.

Unlocking Transfer Learning in Argumentation Mining: A Domain-Independent Modelling Approach

Thiemo Wambsganss[1], Nikolaos Molyndris[1], Matthias Söllner[1,2]

[1] University of St.Gallen (HSG), Institute of Information Management, St.Gallen, Switzerland
{thiemo.wambsganss, matthias.soellner}@unisg.ch, nikolaos.molyndris@student.unisg.ch
[2] University of Kassel, Information Systems and Systems Engineering, Kassel, Germany,
soellner@uni-kassel.de

Abstract. Argument identification is the fundamental block of every Argumentation Mining pipeline, which in turn is a young upcoming field with multiple applications ranging from strategy support to opinion mining and news fact-checking. We developed a model, which is tackling the two biggest practical and academic challenges of the research field today. First, it addresses the lack of corpus-agnostic models and, second, it tackles the problem of human-labor-intensive NLP models being costly to develop. We do that by suggesting and implementing an easy-to-use solution that utilizes the latest advancements in natural language Transfer Learning. The result is a two-fold contribution: A system that delivers state-of-the-art results in multiple corpora and opens up a new way of academic advancement of the field through Transfer Learning. Additionally, it provides the architecture for an easy-to-use tool that can be used for practical applications without the need for domain-specific knowledge.

Keywords: Argumentation Mining, Argument Identification, Transfer Learning, Natural Language Processing

1 Introduction

The identification and classification of argumentation, so-called Argumentation Mining (AM), has received special attention from researchers and practitioners since it enables the automated extraction of structured information from textual sources. The potential of AM has been investigated in different research domains, addressing some of the most challenging multidisciplinary issues in knowledge extraction. Issues such as automated skill learning support [1], accessing argumentation flows in legal texts [2], better understanding of costumer opinions in user-generated comments [3], or fact-checking and de-opinionizing of news [4] have been approached with AM. However, the identification and detection of arguments is not yet used widely in practice, since the practical implementation of AM faces two main challenges in Machine Learning (ML) and Natural Language Processing (NLP): First, even though good modelling results have been achieved in one domain, research on models that perform well for multiple domain corpora is still lacking. Secondly, successful AM use cases usually depend on handcrafted NLP features that require a significant amount of manual labor and deep

15th International Conference on Wirtschaftsinformatik,
March 08-11, 2020, Potsdam, Germany

domain knowledge, which organizations often struggle to provide. As a result, very domain-specific models have been built that reached satisfying results in a respective domain but were mostly useless in different domain corpora and therefore not applicable in a practical environment.

One promising solution avenue is utilizing Transfer Learning for Natural Language Processing. Transfer Learning is a concept in the field of Machine and Deep Learning that tries to transfer the model learnings from one unrelated topic to another, effectively reducing the development time and increasing the prediction power of the models compared to domain-specific development [5]. Therefore, we aim to address those challenges of AM by using a novel Transfer Learning solution that is inter-domain applicable and does not require any labor-intensive NLP features. Current solutions in AM with unsupervised learning (e.g., [6]) or classification approaches using embedding structures and neural networks ([7], [8]) fall short of solving those issues, since they are either not generalizable or very domain-specific. Hence, we aim to contribute to literature and practice by presenting a novel solution that works on a Deep Learning model architecture and enables future scientists and researchers to build AM pipelines without intensive effort. Our solution is based on Deep Transformers for Natural Language Transfer Learning proposed by Delvin et al. [9]. This new approach has been successfully applied in multiple NLP tasks that require language understanding with state-of-the-art results. The solution components adapted to AM that are proposed to solve the aforementioned problems are: First, an easy-to-train and easy-to-use Transfer Learning model to identify arguments and, second, a standardized corpus design to simplify new corpus introductions. To tackle the stated challenges, we develop an artifact following the Cross Industry Standard Process for Data Mining (CRISP-DM) Model [10]. As we described above, we aim to develop a novel modelling approach to identify argumentative text from multiple corpora. The CRISP-DM Model was especially built to develop data modeling approaches like ours. It consists of six different stages, in which a modelling artifact is iteratively developed. To the best of our knowledge, there is no study that developed a solution to the stated AM challenges using end-to-end Transfer Learning approaches.

The remainder of the paper is structured as follows: First, we provide the necessary conceptual background on Argumentation Mining and present the identified challenges based on a systematic literature review following Webster and Watson [11] and Vom Brocke et al. [12]. Next, we present our CRISP-DM methodology in section three and explain the building and evaluation of the model in section four. Finally, we present and evaluate our results, followed by a discussion about the limitations and contributions of our study.

2 Conceptual Background

In the following, we will introduce the reader to the basics of Argumentation Mining, present an overview of the related work on Argumentation Identification and briefly explain the concept of Transfer Learning.

2.1 Argumentation Mining

The foundation of Argumentation Mining is argumentation theory. Argumentation theory is about analyzing the structure and the connection between arguments. One of the most prominent argumentation models is the Toulmin model [13]. Toulmin's model asserts that a "good" argument involves a logical structure built on ground, claim, and warrant, whereas the grounds are the evidence used to prove a claim. Walton et al. developed the so-called "argumentation schemes" that use the Toulmin's type of reasoning [14].

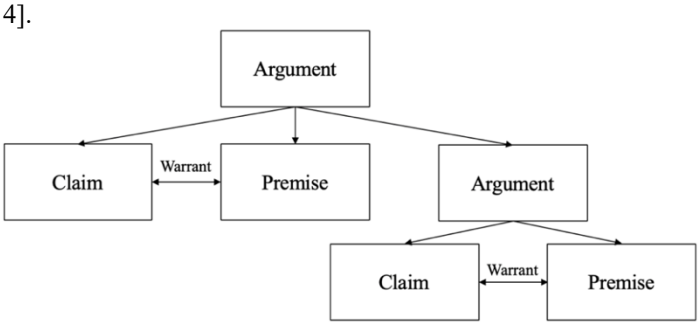

Figure 1. Classic argument tree structure based on Toulmin model [13]

It is commonly considered that "*Claim*", "*Premise*", and "*Warrant*" are the main components of every argument, and the rest are supporting sub-argument parts that may or may not exist in an argument (Figure 1).

Argumentation Mining itself aims to identify these components of an argumentation model with NLP and ML. It falls under the category of computational argumentation, which encompasses a variety of tasks. These tasks include identifying the argumentation style [15], in which arguments are classified as "factual" or "emotional" in order to understand the characteristics better. Identifying the reasoning behind the stance of the author by creating a classifier using the stance classification [16], identifying arguments to be used as summarization pointers [17] , or ranking arguments according to how convincing they are using a joint model with one deep learning module in it [18]. Following Lippi and Torroni 2016 [19], the most related subtasks of Argumentation Mining can be summed up as:

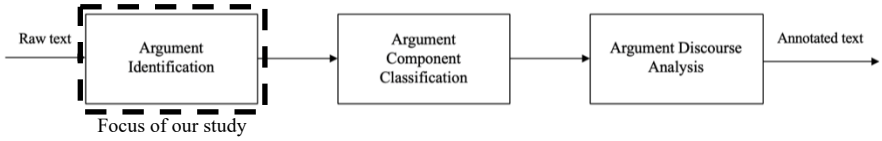

Figure 2. Argumentation Mining pipeline

- **Argument Identification,** which is concerned with identifying the argumentative parts in raw text and setting up its boundaries versus a non-argumentative text.
- **Argument component classification,** which is the subtask of which the primary purpose is to classify the components of the argument structure. Classifying an

argumentative text into claims or premises is one popular way of tackling the target of this subtask.

- **Argumentative discourse analysis**, during this subtask, the researcher tries to identify the discourse relations between the various components existing in the argument. A typical example of this subtask is the identification of whether a support or an attack relationship exists between the claim and the premise.

In our study we are focusing on the challenges of argument identification, since this is usually the first step for an AM architecture and, thus, the foundation of every AM architecture (see Figure 2). Therefore, in order to analyze the current state of literature, potential research challenges and research gaps, we conducted a systematic literature review based on Webster and Watson [11] and Vom Brocke et al. [12]. Details about the methodology of the literature review (such as search strings and data bases) are explained in section 3. We summarize the results of the very review in the following paragraph as related work on Argumentation Identification.

2.2 Related Work on Argument Identification

One of the first advancements in the field of Argumentation Identification came in [2], where Machine Learning techniques were used for the first time to develop an argument identifier in legal texts. They used heavily handcrafted features and relatively simple Machine Learning algorithms, but their results were encouraging. Winkels et al. 2013 [20] were among the first who experimented with unsupervised techniques in Argumentation Mining. Their results showed that pure unsupervised clustering does not yield satisfactory results [21], and Habernal and Gurevych 2015 [22] employed the new idea of word embeddings in a semi-supervised fashion for the argument identification subtask. In some cases, the results yielded a 100% improvement over previous attempts on complex online corpora. In the continuous attempt to limit the number of handcrafted features and corpus-specific knowledge, research has mainly shifted towards Deep Learning. The most common architecture is bidirectional Long-Short Term Memory (LSTM) Neural Networks fed with word embeddings [23, 24, 7]. In general, Deep Learning frameworks tend to give state-of-the-art results, which approach human performance. However, the problem remains that when the model is getting introduced to an entirely new corpus, the accuracy falls significantly [25]. In addition, no effort has been made to generalize the models for new corpora, possibly because of model complexity and a high skill barrier to use [26]. As a result, there is a divide in AM between traditional ML techniques with a lot of manual labor and new DL techniques that require significant skill specialization. This divide is the challenge that we are addressing with our proposed model. In Table 1 we display the most important studies and the AM tasks the respected studies contributed to. It can be seen in Table 1 that the majority of research has been using traditional Machine Learning approaches, which tend to be overly specialized. The two Deep Learning papers that significantly impacted the space with their results are also single-corpus-focused and employ a handcrafted model, which makes it hard to reproduce and use.

Table 1. Representation of the methods and level of analysis in Argumentation Mining, according to the literature

Study	Corpus	Argument Identification	Argument Classification	Discourse Analysis	Method*
[2]	Araucaria	x			NB
[27]	Araucaria ECHR	x	x	x	SVM
[28]	BioNom	x	x		SVM
[1]	ComArg corpus	x			SVM
[21]	Greek news articles	x	x		CRF
[29]	Web discourse	x			SVM, LR, NB
[22]	Web Discourse	x	x		SVM
[30]	Case laws (ECHR)	x			SVM, RF
[23]	Multiple	x			LSTM
[31]	Web annotated text	x	x		LSTM

NB: Naïve Bayes, SVM: Support Vector Machine, CRF: Conditional Random Fields RF: Random Forest, LSTM: Long-Short Term Memory Neural Network based architecture

2.3 Transfer Learning

Unlike traditional supervised and semi-supervised Machine Learning algorithms, which assume that the distribution of the labeled and the unlabeled data is the same, in Transfer Learning there are no assumptions about the distributions, domain, or task, and it allows them to be different from each other in training or testing [32]. Transfer Learning is heavily inspired by the way humans acquire knowledge; learning how to recognize a cat can help in recognizing a tiger or learning Spanish can help with learning French. In general, the study of Transfer Learning is motivated by the fact that people can intelligently apply previous knowledge to solve new problems [5].

For this paper, we are focusing on inductive Transfer Learning (Inductive), where we use a model trained on a completely different corpus to do inference on another corpus for another kind of task [33]. For this kind of transfer learning we are using a type of Recurrent Neural Network called "*Transformer*" [33]. Transformers are improving upon the LSTM-attention mechanism presented in Vaswani et al. 2017 [34] by being able to circumvent the LSTM-attention mechanism problem of being able to process data only sequentially in the encoding step, potentially missing non-sequential information on one side of the sentence [35]. Transformers are able to parallelize the attention mechanism effectively, "*looking*" both before and after the "to-be-predicted" token.

More specifically, in Argumentation Mining, applications of Transfer Learning have been lagging behind. This is in line with the observation that Deep Learning in AM literature has only started to appear in the past couple of years. That being said, while word embeddings as seen in Young et al. 2018 [36] and Mikolov et al. 2006 [37] have been used for quite some time in Argumentation Mining as either features in Machine Learning algorithms or as input layers in Deep Learning models [38, 39, 23], no attempt has been made to use a complete Transfer Learning Pipeline as described above.

3 Methodology

The first step of our research was to identify possible gaps and challenges of Argumentation Mining in scientific literature. Therefore, we have drawn on the approaches by Webster and Watson [11] and Vom Brocke et al. [12]. Based on well-cited literature in AM, such as [18], [25] and [26], we identified different key words, which researchers used to describe the pipelines of AM. Based on these, we built the following search strings trying to incorporate the previously captured namings: ("*Argumentation*" AND "*Mining*"), ("*Argument*" AND "*Identification*"), ("*Argument*" AND "*Classification*").

To find relevant literature, we applied the search string to the following six databases: *AISeL, ACM Digital Library, EBSCO, IEEE Explore, ProQuest ABI Inform and Science Direct*. Table 2 shows the hits and the relevant papers of each database.

The database search resulted in 12,529 hits. Titles, abstracts, and keywords were screened to fit the abovementioned definition of Argumentation Mining and the application of NLP and ML to the scope of our study. We excluded papers that did not refer to a technical perspective on Argumentation Mining. Multiple papers were excluded due to a different research scope described in their abstract, e.g., several papers talked about argumentation in different domains, e.g., in learning science or mineral mining. Moreover, numerous studies were excluded since they were talking about classification or identification of information in completely different domains, e.g., topic classification of user-generated content or identification of user needs on Twitter.

Table 2. Overview of found hits and relevant papers for each database.

	Databases											
	AISeL		**ACM**		**EBSCO**		**IEEEXplore**		**ProQuest**		**ScienceDirect**	
Search strings	**Hits**	**Relevant**	**Hits**	**Relevant**	**Hits**	**Relevant**	**Hits**	**Relevant**	**Hits**	**Relevant**	**Hits**	**Relevant**
"Argumentation" AND "Mining"	206	5	103	7	8	0	26	12	3446	3	206	5
"Argument" AND "Identification"	369	2	126	4	8	0	73	9	50	4	862	0
"Argument" AND "Classification"	269	3	77	3	9	0	209	7	5420	3	1062	0
Total number of relevant literature selected from 67 screened papers	With Duplicates:				**67**	+	12 Forward Search				=	**68**
	Without Duplicates:				**50**		6 Backward Search					

This screening process resulted in 67 potentially relevant papers which conducted studies on Argumentation Mining, Argumentation Identification or Argumentation Classification. After the elimination of all duplicates, we had 50 relevant papers left.

Afterwards, forward and backward search was carried out according to vom Brocke et al. [11]. Through screening the references, 18 studies were added to the list, resulting in 68 relevant papers, which were the basis for our related work section in the theoretical background section. We found two main literature gaps, namely, the lack of corpus-agnostic models and problem of human-labor-intensive NLP models being costly to develop.

Next, we formulate our hypothesis that Transfer Learning might provide a solution to both challenges. In order to prove our hypothesis, we aim to develop a new Argumentation Identification pipeline based on the current state of Transfer Learning. Afterwards, we want to evaluate the results of our pipeline for different corpora compared with the latest results published in literature. In order to do so, we develop an artifact following the Cross Industry Standard Process for Data Mining (CRISP-DM) Model, which is illustrated in Figure 3 [10]. The model describes a standardized approach for Data Mining problems from a practical point of view, followed by the data understanding, the data preparation, and the data modelling.

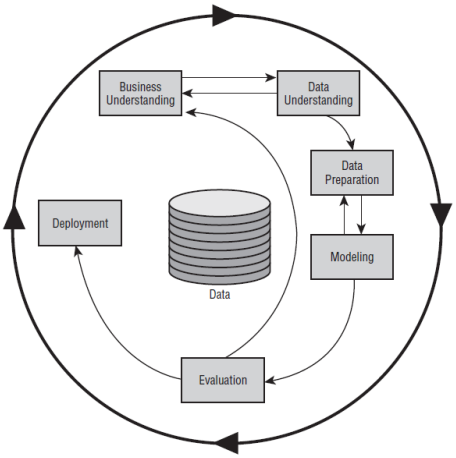

Figure 3. Cross Industry Standard Process for Data Mining (CRISP-DM) Model [10]

Our approach is divided into the five iterative stages (excluding the deployment stage). In the first stage, we analyzed the current state of Argumentation Identification achievements in literature. Second, we investigate different corpora and their results at the current state in Argument Identification in terms of precision, recall, or accuracy across multiple domains. In fact, the goal of our research is to find a model that is domain-agnostic; however, we did not find any model in literature that achieved this goal. Third, we build an artifact that is able to identify an argument in an unknown corpus. This is achieved by using Natural Language Processing and Deep Learning algorithms to classify a text piece as an argument or a non-argument. The fourth stage is an iterative process of evaluation and revision of the model based on various performance metrics such as the f1-score. In this stage, we expand the model usefulness and applicability by adding additional corpora and by identifying the best

hyperparameters for our model based on the results. Finally, in the fifth stage, we draw conclusions based on the iterative process and the results.

Our approach is developed using the programming language Python 3.7 for the ML applications, since it is widely known, easy to use, and supports major libraries for NLP and ML tasks. The ML-related algorithms are called from the Google-supported tool Scikit-learn [40] and its major ML packages [41]. For DL, TensorFlow and its integrated Keras [42] are called.

4 Implementation

The goal of the research is to create a model that augments the current Argumentation Mining techniques. Specifically, it aspires to reveal a unified model architecture that is corpus-agnostic and reduces the manual corpus-specific work. In order to accomplish this, we propose a Transfer Learning approach based on Deep Bidirectional Transformers for Language Understanding (BERT) as seen in [9]. We hypothesize that this model architecture is uniquely suited to act as a unified, domain-agnostic, and easy-to-develop solution for AM, due to its architectural novelties outlined in section 4.4. In order to validate the hypothesis, we conduct research structured as a CRISP-DM cycle, as it is demonstrated above. The first phase, business understanding, is explained in the introduction and the theoretical background of this work.

4.1 Data Understanding: Corpora Collection

The solution to the stated problems begins with the finding of a representative corpus for Argumentation Mining, followed by the construction of the model, and the training and evaluation of the classification algorithms. The first dataset acted as a blueprint for the model. The model specifications were built according to the first dataset to reduce the noise of the different annotations and assumptions of different corpora. The blueprint corpus had to be a well-structured, multi-used corpus that would minimize the possibility of corpus-specific irregularities and allow the shift of tuning and research towards the architecture and model themselves. After a thorough search of the available datasets, it became apparent that the most suitable and used corpus is the Student Essays corpus from [39], containing 402 annotated student essays. The corpus format has been used multiple times, the text nature (student essays) provides the best representation of a user trying to argue in a structured way. The corpus is divided into essays, each one has embedded annotated argument components "*Major Claim*", "*Claim*", "*Premise*". In order to deduct a binary-labelled corpus from this, the labels above are transformed into the labels "*Argument*" and "*Non-Argument".* The labels "Major *Claim*", "*Claim*", and "*Premise*" are labelled as "*Argument*", and the rest of the text is labelled as "*Non-Argument*". However, to achieve our goal to build a domain-crossing argumentation identification model, we needed to find additional corpora from different domains. During our systematic literature review, we decided to include the following corpora from different domains for our model:

- *AracauriaDB*

This corpus was introduced by Reed et al. 2008 [44] and includes various sources, such as parliamentary records, news, and court summaries. The annotation quality is unknown, but it is among the very first and the most used corpora in Argumentation Mining.

- *User-generated web discourse*
  This corpus consists of blog posts and user comments on various issues that are annotated with claims, premises, backings, and rebuttals according to the Toulmin model [13].
- *Wikipedia Blog comments*
  Biran and Rambow 2011 [45] annotated two datasets. One was blog posts and their comments from LiveJournal and the second one was comments on the Wikipedia talk (WT) pages, where discussions on Wikipedia entries are made. We chose to incorporate the WT corpus since the quality is significantly better.
- *Combination*
  Following the standardization procedure of the proposed pipeline, all corpora have the same structure and labelling scheme. Thus, the combination of all of them was introduced to the system with the rationale of getting a holistically trained model.

In Table 3 we outline the results of the corpora's stated metrics in the study of Argumentation Identification at the current state.

Table 3. State-of-the-art results of binary classification based on current literature

Study	Corpus	Size	Method	Metric	Result
[39]	Student essays	402 Essays	CRF	f1	85.40%
[2]	AraucariaDB	~2800 sentences	MNB	Accuracy	73.75%
[38]	UGC Web Discourse	3900 sentences	SVM	f1	71.40%
[25]	Wikipedia Blog Comments	1985 documents	CNN	f1	60.50%

MNB: Multinomial Naïve Bayes Classifier, SVM: Support Vector Machine, CRF: Conditional Random Fields, CNN: Convolutional Neural Network

4.2 Data Preparation

The data preparation is split into two parts, the data standardization and the data preprocessing. The data standardization is a meta-architecture that makes sure that each corpus abides by the same rules before entering the model. Data preprocessing is the steps that have to be taken for the corpus to be understood by the model's architecture. Since different corpora have completely different annotation schemes, in order to homogenize but also to keep tokens at the above-word length, the sub-sentence length was chosen. Each sentence was split into sub-sentences signaled by commas (',') or other punctuation ('?','!',';') and each token was assigned the "parent" label of "Argument" or "Non-Argument". All extra whitespaces, punctuation, and special

characters were removed. Subsequently, since BERT requires some unique text preprocessing steps in order to work, the steps below had to be taken:

1. The labels should be converted to 0 or 1.
2. The text should be transformed to lower case.
3. The text input must be padded to be a multiple of the batch size in order to be run on the Google Cloud TPU without errors.
4. The text should be split into special sub-word tokens.
5. The text has to be converted into input features in order to be understood by the Deep Learning framework.

Except for converting the labels into binary, the preprocessing steps are handled by the tokenizer of BERT, minimally adapted to work on a custom classification scenario. Finally, since the classes on most corpora are very imbalanced, a balancing method was chosen. Taking into consideration that the data size is small (and the range of options that are available to avoid overfitting), the oversampling technique from the Sklearn python library was used on the training sets.

4.3 Modelling

After the meta-modelling and the preprocessing of the text, the result is inserted into the proposed learning architecture.

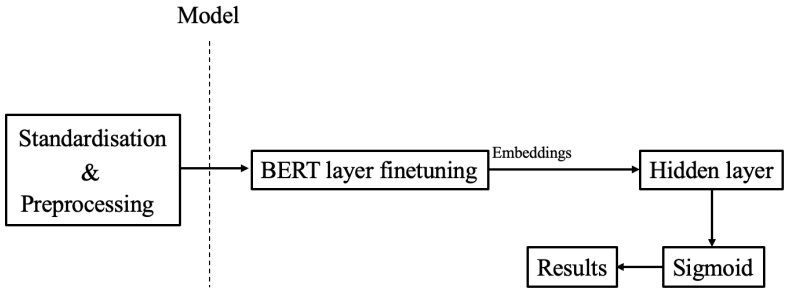

Figure 4. Outline of our proposed learning architecture

The goal of this model is to be a flexible, corpus-agnostic, and accurate way of identifying arguments that would be a foundation piece for the process of building an automated argument feedback system. BERT is the perfect candidate for this because of its flexibility and generalization. BERT is a pre-trained deep learning model that has been trained unsupervised using the whole Wikipedia corpus. The novelty of this architecture and what makes it ideal for Transfer Learning tasks is that it is able to capture semantic information from text, which can then be used for other downstream tasks without the need for retraining. BERT is able to capture this semantic information by implementing two architectural novelties. The first one breaks each word into sub-word tokens that are able to much better cope with unique words and misspellings. The second one predicts a token by looking both at the text that precedes it and the text that follows using “Transformers” instead of the traditional LSTM approach as seen in [36,

46]. Practically, BERT is available through Tensorflow Hub[1], a server that hosts multiple Google-released models that can be called upon with simple API requests and automatically run on Google Cloud TPUs by default. The API access is done through Python and with the libraries of Tensorflow and Keras. The proposed architecture consists of a corpus-agnostic BERT implementation with ten fine-tuning layers with an additional single hidden layer. Fine-tuning enables to inexpensively train some of the model's parameters, making the model specialized to Argumentation Mining while keeping the knowledge that the model acquired multiple days of expensive training. The last hidden layer is a Recurrent Neural Network with 512 nodes that takes the BERT output and learns to feed into a sigmoid layer that classifies each input into one of the two classes.

Since almost all of the corpora are small, regularization techniques are needed to avoid overfitting and to stabilize the model. A dropout layer with a dropout rate of 0.7 was introduced. The whole model is considered to be quite "deep"; hence, it tends to overfit quite fast with small datasets. Thus, a high dropout rate ensures that this does not happen as fast. Also, the "callback" Keras feature was used to save the best performing model, to decrease the learning rate once the learning has plateaued, and finally to stop the training once the evaluation metric has stopped increasing. Practically, this corresponds to the introduction of the Keras methods "ModelCheckpoint", "EarlyStopping", and "ReduceLROnPLateu". Since the goal is to determine a corpus-agnostic model, the objective of the hyperparameter tuning is to find the parameters that work best for all used corpora. A Grid Search for all corpora and all parameters would be too time-consuming, so the Student Essays corpus was used as a blueprint to significantly limit the universe of possible hyperparameters. Here, the following hyperparameters were optimized with a "*sparse*" grid-search: *number of fine-tuned BERT layers, number of layers in the RNN, number of nodes per Layer, regularization techniques, optimizer, loss* and *batch size*. This search resulted in choosing the "*adam*" optimizer, the "*binary crossentropy*" loss, and the introduction of a "*dropout layer*" as a best practice technique. After performing a Grid Search of the student essays, the possible space was substantially reduced to the number of fine-tuned layers, since it was the main parameter with by far the most significant performance variability. Subsequently, a Grid Search of the layers off all corpora was done to reveal the optimal parameters.

4.4 Evaluation

After each training, the algorithm is used to classify the posts in the previously unseen test set. The classification results (positive/negative) are compared with the true labels. The percentage of correctly classified posts compared to the total number of posts is called accuracy. Precision is the fraction of the truly positive posts among all positively classified posts, and recall is the fraction of truly positive posts that have been classified as positive. The f1-score balances between precision and recall and thus presents the most suitable criterion for our use case, since precision and recall are both equally

[1] https://tfhub.dev/

important. However, we have recorded accuracy, recall, precision, and f1-score for all our classification experiments. The results are stated in Table 4.

Table 4. Achieved results for different domain corpora with the same BERT learning model

Corpus	Accuracy	Precision	Recall	f1
AracauriaDB	97.46%	95.51%	92.37%	**93.10%**
Student Essays	80.00%	80.41%	91.16%	85.19%
Web Discourse	80.00%	81.49%	87.65%	**84.15%**
Blog comments	66.59%	70.00%	80.75%	**74.86%**
Combined	75.79%	73.22%	80.70%	**76.57%**

Comparing our results achieved with one unique model (Table 4) with the results for these corpora at the current state (Table 3) for Argumentation Identification, one can observe that our proposed model produces state-of-the-art results for almost all datasets except the student essays. The difference is minimal, and this signifies the miniscule trade-off (0.21%) between a highly specialized and complex-to-develop model versus the proposed generalized solution. It is hypothesized that AracauriaDB has the best score because it is the easiest of all; even a simple Machine Learning system [2] achieves a score of ~74% accuracy. On the other end of the spectrum, it is hypothesized that the Blog comments corpus has the worst score because the nature of the corpus is much less clean, and it is difficult to analyze and deduct argumentation. As a matter of fact, in Daxenberger et al. 2017 [25] even Deep Learning methods fail to distinguish a claim from a non-claim in an above-chance probability.

Finally, to the best of our knowledge, there is no other research that tried to create a generalized model out of multiple AM corpora. The combination of the corpora was made possible by creating a unified data standardization pipeline that allowed differently annotated corpora to be concatenated into a single, multi-domain corpus. The result demonstrates that even a BERT module pre-trained on non-argumentative text (Wikipedia) can produce very good results that imply language understanding at a multi-domain corpus that is structurally and semantically radically different both from the pre-trained model and from each of the sub-corpora.

5 Discussion

The aim of the study was to solve current challenges in Argumentation Mining by utilizing a novel Transfer Learning solution that is inter-domain applicable and does not require any labor-intensive NLP features. We conducted a systematic literature review based on Webster and Watson [11] and Vom Brocke et al. [12] to analyze the current state and potential gaps of Argumentation Mining literature. Our hypothesis was that new advantages of transfer learning can solve the challenges of domain-agnostic models and labor-intensive modelling approaches. We developed a new modeling approach utilizing BERT, and we prove that this is a suitable technology avenue for those challenges.

Our contribution is twofold: First, we contribute to scientific literature by demonstrating a new modelling approach and solution to current problems of Argumentation Mining. Our proposed model demonstrates that a whole new path has opened up for Argumentation Mining. Utilizing the power of Transfer Learning models such as BERT provides a potential solution to the fragmentation of the Argumentation Mining research into multiple domain-specific nodes and into a Machine Learning vs Deep Learning dichotomy. Our research is the first step in implementing such techniques in the whole Argumentation Mining pipeline.

Second, we contribute to practice by providing a use case on how to utilize AM in interdomain applications without intensive human modelling. Even though the proposed system uses state-of-the-art technology to be able to infer meaning from pools of text, the availability of the toolset in an open-source and third-party maintained fashion is making it accessible to practitioners that had minimal exposure to deep learning before. In addition, the API access minimizes the time costs of running such a deep model since it is run remotely by default. The reduced time costs allow for much easier experimenting and even multiple practical applications with minimal training data input or even direct inferencing from related corpora. In addition, we open-source the whole pipeline in a repository. We provide a potential application model for organizations and thus assist them to leverage the identification and detection of arguments in practice.

The model is based on the best structured available datasets. Of course, in real world scenarios the textual data is less structured than the one encountered. For future work we see two possible paths of immediate action: 1) To expand the model to argument classification and discourse analysis 2) To expand the BERT model's capabilities by pre-training the whole model on AM-related data instead of the generic Wikipedia data that was used here.

6 Conclusion

We have presented a novel approach for identifying argumentative text in a corpus that is domain-agnostic, produces state-of-the-art results in multiple corpora, and is significantly easier to develop for existing solutions. First, we introduced a new model based on state-of-the-art Transfer Learning architecture. Second, we implemented a standardized approach to the data preparation and the preprocessing of four different corpora and their combination. Third, we applied the model to multi-domain corpora, achieving state-of-the-art results in most of them while using high-level APIs available for future work. All in all, we believe that this paper has the potential to open up a new, unified approach to AM, utilizing the advancements of Transfer Learning in Natural Language Processing, effectively bypassing the chronic issue of small-sized corpora in AM.

References

1. Stab, C., Gurevych, I.: Identifying Argumentative Discourse Structures in Persuasive Essays. 46–56 (2014).
2. Moens, M., Boiy, E., Reed, C.: Automatic Detection of Arguments in Legal Texts. (2007).
3. Boltuži, F., Šnajder, J.: Back Up Your Stance : Recognizing Arguments in Online Discussions. 1–43 (2014).
4. Dusmanu, M., Cabrio, E., Villata, S.: Argument Mining on Twitter: Arguments, Facts and Sources. 2317–2322 (2018).
5. Pan, S.J., Yang, Q.: A Survey on Transfer Learning. IEEE Trans. Knowl. Data Eng. 1–15 (2009).
6. Ferrara, A., Montanelli, S., Petasis, G.: Unsupervised Detection of Argumentative Units though Topic Modeling Techniques. 97–107 (2018).
7. Gema, A.P., Winton, S., David, T., Derwin, S., Muhsin, S., Wikaria, G.: It Takes Two To Tango: Modification of Siamese Long Short Term Memory Network Attention Mechanism in Recognizing Argumentative Relations in Persuasive Essay. Comput. Intell. 449–459 (2017).
8. Xu, J., Yao, L., Li, L.: Argumentation based joint learning: A novel ensemble learning approach. PLoS One. 10, 1–21 (2015).
9. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (2018).
10. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Daimlerchrysler, T.R., Shearer, C., Daimlerchrysler, R.W.: Step-by-step data mining guide. SPSS inc. 78, 1–78 (2000).
11. Webster, J., Watson, R.T.: Analyzing the past to prepare for the future : Writing a literature review Reproduced with permission of the copyright owner . Further reproduction prohibited without permission . MIS Q. (2002).
12. Vom Brocke, J., Simons, A., Riemer, K., Niehaves, B., Plattfaut, R., Cleven, A.: Standing on the shoulders of giants: Challenges and recommendations of literature search in information systems research. Commun. Assoc. Inf. Syst. (2015).
13. Toulmin, S.E.: Introduction to Reasoning. (1984).
14. Walton, D., Reed, C., Macagno, F.: Argumentation Schemes. Cambridge University Press, Cambridge (2008).
15. Oraby, S., Reed, L., Compton, R.: And That's A Fact: Distinguishing Factual and Emotional Argumentation in Online Dialogue. (2000).
16. 16. Hasan, K.S., Ng, V.: Why are You Taking this Stance ? Identifying and Classifying Reasons in Ideological Debates. 751–762 (2014).
17. Richard, A., Suhartono, D., Chowanda, A.D., Setiadi, C.J., Jessica, C.: Automatic Debate Text Summarization in Online Debate Forum. Procedia Comput. Sci. 116, 11–19 (2017).
18. Habernal, I., Gurevych, I.: Which argument is more convincing ? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM. 1589–1599 (2016).
19. Lippi, M., Torroni, P.: Argumentation Mining : State of the Art and Emerging Trends. 16, (2016).
20. Winkels, R., Douw, J., Veldhoen, S.: Experiments in automated support for argument reconstruction. ICAIL. (2013).
21. Sardianos, C., Katakis, I.M., Petasis, G., Karkaletsis, V.: Argument Extraction from News. 56–66 (2015).
22. Habernal, I., Gurevych, I.: Exploiting Debate Portals for Semi-Supervised Argumentation Mining in User-Generated Web Discourse. 2127–2137 (2015).

23. Schulz, C., Eger, S., Daxenberger, J., Kahse, T., Gurevych, I.: Multi-Task Learning for Argumentation Mining in Low-Resource Settings. (2017).
24. Rajendran, G., Chitturi, B., Poornachandran, P.: Stance-In-Depth Approach Deep Neural Approach to to Stance Stance Classification. Procedia Comput. Sci. 132, 1646–1653 (2018).
25. Daxenberger, J., Eger, S., Habernal, I., Stab, C., Gurevych, I.: What is the Essence of a Claim ? Cross-Domain Claim Identification. EMNLP. (2017).
26. Lippi, M., Torroni, P.: Argument mining: A machine learning perspective. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2015).
27. Palau, R.M., Moens, M.: Argumentation Mining : The Detection, Classification and Structure of Arguments in Text. ICAIL. (2009).
28. Ozyurt, I.B.: Automatic Identification and Classification of Noun Argument Structures in Biomedical Literature. 9, 1639–1648 (2012).
29. Zhang, G., Purao, S., Zhou, Y., Xu, H.: Argument detection in online discussion: A theory based approach. AMCIS Proc. 1–10 (2016).
30. Poudyal, P., Goncalves, T., Quaresma, P.: Experiments On Identification of Argumentative Sentences. 398–403 (2016).
31. Stab, C., Miller, T., Gurevych, I.: Cross-topic Argument Mining from Heterogeneous Sources Using Attention-based Neural Networks. (2018).
32. Torrey, L., Shavlik, J.: Transfer Learning. Encycl. Sci. Learn. 3337–3337 (2012).
33. Jordan, M.I., Mitchell, T.M.: Machine learning: Trends, perspectives, and prospects, (2015).
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need. (2017).
35. Bahdanau, D., Cho, K., Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate. 1–15 (2014).
36. Young, T., Hazarika, D., Poria, S., Cambria, E.: Recent trends in deep learning based natural language processing, (2018).
37. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. CrossRef List. 1, 1–9 (2006).
38. Pennington, J., Socher, R., Manning, C.: Glove: Global Vectors for Word Representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2014).
39. Habernal, I., Gurevych, I.: Argumentation Mining in user-generated web discourse. Comput. Linguist. (2017).
40. Stab, C., Gurevych, I.: Parsing Argumentation Structures in Persuasive Essays. (2017).
41. Varoquaux, G., Buitinck, L., Louppe, G., Grisel, O., Pedregosa, F., Mueller, A.: Scikit-learn. GetMobile Mob. Comput. Commun. (2017).
42. Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., Vanderplas, J., Joly, A., Holt, B., Varoquaux, G.: API design for machine learning software: experiences from the scikit-learn project. 1–15 (2013).
43. Chollet, F.: Keras Documentation, (2015).
44. Reed, C., Palau, R.M., Rowe, G., Moens, M.-F.: Language Resources for Studying Argument. In: Proceedings of the 6th conference on language resources and evaluation (2008).
45. Biran, O., Rambow, O.: Identifying justifications in written dialogs. In: Proceedings - 5th IEEE International Conference on Semantic Computing, ICSC 2011 (2011).
46. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. (2018).