# Hybrid Attention-Based Deep Claim Identification from Argumentative Text

Kilol Gupta
Columbia University
kilol.gupta@columbia.edu

Tariq Ahlindi
Columbia University
tariq@cs.columbia.edu

Tuhin Chakrabarty
Columbia University
tc2896@columbia.edu

## Abstract

*Claim identification (a sub-task of argument mining) is an important task for many applications such as fake-news detection, political and legal affairs, educational purposes (e.g. auto grading of essays) etc. Success of current approaches is limited to situations where training and test data are from the same domain. In this paper, we have explored 3 deep learning models for the task of in-domain and cross-domain claim identification. We proceed to empirically prove that higher performance can be achieved for this task using word level attention applied to a bidirectional long short term memory (BiLSTM) network augmented with word embeddings which are learned from argument lexicon.*

## 1. Introduction

The claim is the principle component of an argument. Different theories had varying definition of what a claim is. One recent definition of a claim is 'a statement that is in dispute and that we are trying to support with reasons' [16]. Argument mining is the process of computationally identifying arguments from text. It starts with separating argumentative units from non-argumentative ones. Then, classifying argumentative units to claims among other types.

Claims are identified with high precision in argumentative discourse but that precision diminishes greatly in less homogeneous datasets. An earlier study found that despite difference in properties of claims across datasets, there are nevertheless properties that are consistent at the lexical level [3]. They conducted many experiments using models with linguistically motivated features and using deep neural networks and found that the choice of training data is crucial when the target domain is unknown. We want to build on their work by exploring other neural network architectures and combining them with linguistic features to improve the performance.

## 2. Literature Review

Existing approaches to argumentation mining take two forms: 1.Multi-document approaches which recognize claim and evidence across multiple documents as described in the works of [6] and [10]. 2.Discourse level approaches which recognize argumentative structure within a single document as described in the following works. [8] use SVM and a hand-crafted context free grammar (CFG) to recognize claims and premises in legal document (domain-specific). [9] adopt a minimum spanning tree (MST) approach for recognizing claims in English 'microtexts'. [14] identify claims in student essays experimenting with several classifiers with best reported performance using Support Vector Machines (SVM) with structural, lexical, syntactic, indicator and contextual features. Taking motivation from this and from [3]'s work using feature engineering, our future experiments will explore a hybrid neural network framework that leverages from both deep learning and handcrafted features .

All of the above discussed work achieve promising results in particular domains but their ability to generalize over cross-domains/heterogeneous texts remain unanswered. [3] in their work tackle the problem of cross-domain claim identification on a discourse level, with each sentence labelled as claim or non-claim. This is because this was the only way to make all data sets compatible to each other.
There has been work on cross-domain claim identification as well. [11] in their work detect claims in LiveJournal blog articles and Wikipedia discussions, but both of these domains are of similar genre i.e. social-media and one can expect similar type of content in them. They adopted self-annotation leading to identical notion of claim. [1] improved argumentation mining using distant supervision but they addressed a different task which was of identification of argumentative sentences. [3] in their work experiment to learn universal feature sets or classifiers that can perform reasonably well across varying source and target domains. Our work is targeted to explore and build deep learning models which are particularly optimized for the task of cross-domain claim identification task. And

hence is different from the work of our reference paper where they are trying to establish the difference between notions of claims across various domains.

From the deep learning architecture point of view, [17] in their work apply hierarchical attention (sentence level and word level) networks to the application of document classification. This paper provided us valuable insight on the working of attention mechanism. But since the claim-identification task has been modeled at the sentence level, we have explored the application of word-level attention for our preliminary set of experiments. [4] in their work attempt to identify claim and premises as individual components along with their relations and links to each other. They formulate argument mining as a sequence tagging problem using the state of the art BiLSTM CNN CRF taggers introduced by [7]. They also formulate argument mining as a dependency parsing problem and argument mining as a multi-task learning and present a comparative analysis of the results using the persuasive essays dataset.

Argumentation mining (AM) finds applications in varied fields such as legal decision making [8], research and analysis on scientific (student) papers as shown in the work by [4] and in document summarization. The importance of argumentation mining in educational domain has been highlighted by the recent works of [5] on assisted writing and scoring of student essays [13]. Claim forms an important element of an argument structure and hence identifying claims from an argument is a compelling problem to solve.

## 3. Problem Formulation

We are using the same datasets used in [3] where we classify claims at the sentence level. We are doing a binary classification of sentences where each sentence is classified as claim if it has at least one token labeled as claim and classified as non-claim otherwise. Below is the list of the six datasets:

1. Microtexts (MT) : *MT Data Source* This dataset consists of German microtexts (MT) of controlled linguistic and rhetoric complexity. Each document includes a single argument and does not exceed five argument components. The scheme models the argument structure and distinguishes between premises and claims, among other properties (such as proponent/opponent or normal/example) [3].
2. Web Discourse (WD) : *WD Data Source* This dataset includes user-generated web discourse such as blog posts, or user comments annotated with claims and premises as well as backings, rebuttals and refutations [3].
3. Persuasive Essay (PE): *PE Data Source* This dataset includes 402 student essays. The scheme comprises major claims, claims and premises at the clause level [3].
4. LiveJournal (OC) : *OC Data Source* This dataset contains annotated claims and premises in online comments (OC) from blog threads of LiveJournal [3].
5. Wikipedia Talk Pages (WTP) : WTP Data Source This dataset consists of annotated documents from Wikipedia Talk Pages with 118 threads.
6. Various Genres (VG): *VG Data Source* The AraucariaDB comprises of various genres (hence the name) such as newspaper editorials, parliamentary records, or judicial summaries. The annotation scheme structures arguments as trees and distinguishes between claims and premises at the clause level. Although the reliability of the annotations is unknown, the corpus has been extensively used in argument mining[3].

MT and PE are monologic in nature while WTP, OC, VG and WD are dialogic in nature.

Table 1. Overview of the corpora

| Datasets | No. of Docs | No. of Tokens | No. of Sentences | No. of Claims |
|---|---|---|---|---|
| VG | 507 | 60,383 | 2,842 | 563 |
| WD | 340 | 84,817 | 3,899 | 211 |
| PE | 402 | 147,271 | 7,116 | 2,108 |
| OC | 2,805 | 125,677 | 8,946 | 703 |
| WTP | 1,985 | 189,140 | 9,140 | 1,138 |
| MT | 112 | 8,865 | 449 | 119 |

**Model Intuition**: The best choice for modelling was Recurrent Neural Networks (RNN). An RNN processes an input sequentially, in a way that resembles how humans do it. It performs the same operation, $h_t = f_W(x_t, h_{t-1})$, on every element of a sequence, where $h_t$ is the hidden state a time step t, and W the weights of the network. The hidden state at each time step depends on the previous hidden states. This is why the order of the elements (words) is important. This process also enables RNNs to handle inputs of variable lengths.

RNNs are difficult to train because gradients may grow or decay exponentially over long sequences. A way to overcome these problems is by using one of the more sophisticated variants of the regular RNN, the Long Short Term Memory (LSTM) network. It introduces a gating mechanism, ensuring proper gradient propagation through the network. For BiLSTM the future input information is reachable from the current state. The basic idea of a BiLSTM network is to connect two hidden layers of opposite direc-

tions to the same output. By this structure, the output layer can get information from the past and future states.

## 4. Methods

We want to investigate whether the different conceptualizations of claims can be assessed empirically and if so, how they could be dealt with in practice. Simply put, the task we are trying to solve is: given a sentence from argumentative text, classify whether or not it contains a claim. We opted to model the claim identification task on sentence level, as this is the only way to make all data sets compatible to each other.
Every document in the 6 data sets is pre-processed using XML DOM parser for reading the data in XML format, documents are segmented into sentences using NLTK sentence tokenizer and then every sentence is annotated as claim, if one or more tokens within the sentence were labeled as claim. Analogously, each sentence is annotated as non-claim, if none of its tokens were labeled as claim. Below is an example of a document from the WTP dataset:

<text><**justification**>Einstein created the theory long before 1933,</**justification**>so <**claim** >it cannot possibly be an American invention.</**claim** >.
Shutz says that "Einstein invented relativity", so adding the category "Swiss inventions" would be appropriate for special relativity, whereas "Austrian and German inventions" would be appropriate for general relativity.</text>

The above document is tokenized into two sentences, the first one is a Claim and the second one is a Non-Claim .

Einstein created the theory long before 1933, so it cannot possibly be an American invention.
**CLAIM**
Shutz says that "Einstein invented relativity", so adding the category "Swiss inventions" would be appropriate for special relativity, whereas "Austrian and German inventions" would be appropriate for general relativity.
**NON-CLAIM**

We also perform some data cleaning manually and ignore the sentences of length less than 5 and sentences which are hyperlinks as they clearly don't fit in an ideal scenario and obviously come as non-claims. The in-domain experiments were carried out in a 10-fold cross-validation setup with fixed splits into training (training + validation) and test data. As for the cross domain experiments, we train on the entire data of the source domain and test on the entire data of the target domain.

The ratio of train, test and dev data sets for in-domain experiments are 8:1:1 and ratio of train and test data sets for cross domain experiments are 9:1.

To address class-imbalance in our datasets, we down-sample the negative class (non-claim) both in the in-domain and cross-domain experiments, so that the positive and negative classes occur approximately in a 1:1 ratio in the training data. Since this means that we discard a lot of useful information (many negative instances), we repeat this procedure 20 times, in each case randomly discarding instances of the negative class such that the required ratio is obtained. So now instead of 1 training file during every cross validation (CV), we have 20 training files. This means that we train 20 models for every cross validation. At the time of testing, we use the majority prediction of this ensemble of 20 trained models for every CV and report F1 scores. We have experimented with three models overall:

1. Bidirectional Long Short Term Memory (BiLSTM) + Dropout (0.25)
2. BiLSTM + Word-Level Attention
3. BiLSTM + Word-Level Attention + Argumentation Lexicon (ArgLex) Features

This is the pipeline diagram for our methods

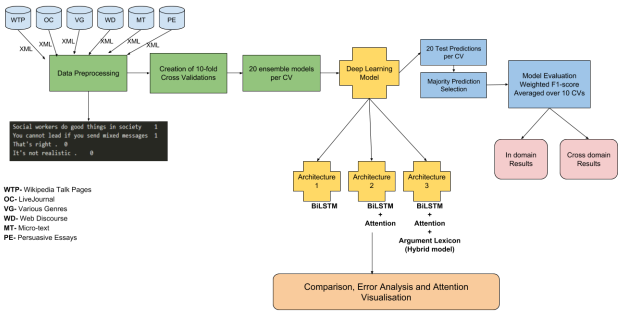


Figure 1: Project Pipeline

We detail each of these layers in the below paragraphs. The models are incremental and each model is built on the previous one.

### 4.1. BiLSTM + Dropout layer

The baseline model consists of a single layer bidirectional LSTM (BiLSTM) .

**Embedding Layer**. The input to the network is a sentence, treated as a sequence of words. We use an embedding layer to project the words: $X = (x_1, x_2, ..., x_T)$

to a low-dimensional vector space $R^E$, where E is the size of the embedding layer and T is the number of words in a sentence. We initialize the weights of the embedding layer with random numbers.
We use bidirectional LSTM (BiLSTM) in order to get word annotations that summarize the information from both directions. A bidirectional LSTM consists of a forward LSTM $\overrightarrow{f}$ that reads the sentence from $x_1$ to $x_T$ and a backward LSTM $\overleftarrow{f}$ that reads the sentence from $x_T$ to $x_1$. We obtain the final annotation for a given word $x_i$, by concatenating the annotations from both directions.

**Output Layer** - We use the representation 'r' as a feature vector for classification and we feed it to a final fully-connected softmax layer which outputs a probability distribution over 2 classes.

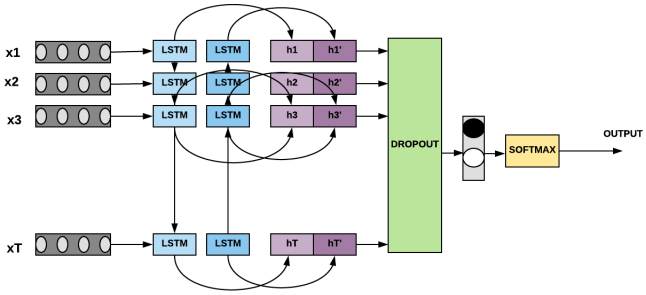


Figure 2: Architecture diagram for BiLSTM + Dropout

### 4.2. BiLSTM + Attention

To the above model, we add an attention layer.
**Attention Layer** - An RNN updates its hidden state $h_i$ as it processes a sequence and at the end, the hidden state holds a summary of all the processed information. Not all words contribute equally to the representation of the meaning of the sentence. In order to amplify the contribution of important words in the final representation, we use an attention mechanism. We aggregate the representation of those informative words to form a sentence vector.

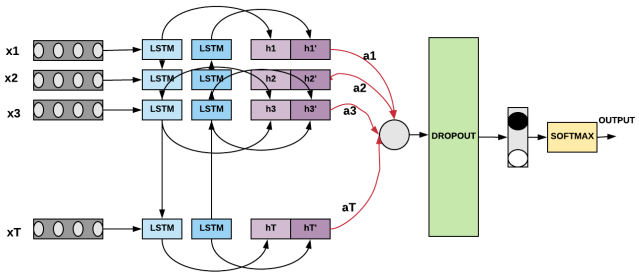


Figure 3: Architecture diagram for BILSTM with attention

### 4.3. BiLSTM + Attention + ArgLex Embeddings

[15] introduced the Arguing Lexicon [1]. The lexicon includes patterns that represent arguing. We take these lexicons and convert them to vectors by looking up the word2vec pre-trained embeddings. Each sentence is then converted to a list of argument lexicon feature vectors. For sentences which do not have an argument lexicon, the list contains a vector of random numbers denoting UNK (unknown). This argument lexicon embedding layer is merged to the output from the attention layer. The merged layer is passed to a fully connected dense layer before connecting to the softmax layer which outputs a probability distribution over the 2 classes (claim and non-claim).

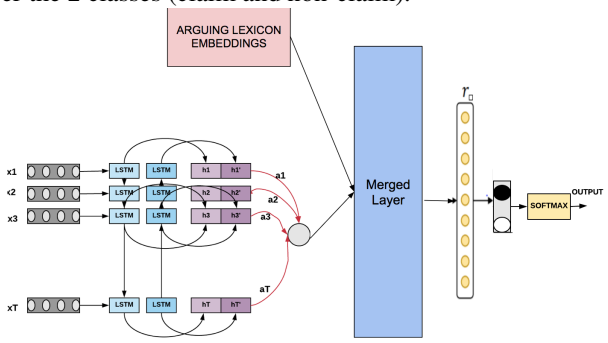


Figure 4: Architecture diagram for hybrid model

## 5. Visualizing attention weights

Words highlighted in deeper colors have greater attention weights. The darker words also interestingly explain the intuition of why they were classified as claims by our deep neural network. Below are a few example sentences predicted as claim by our neural network architecture and the associated relative attention weight.

Ex.1: *It's certainly true that this article shouldn't contain full detail about the issue, as it has its own article*


Figure 5: Sentence from WTP

Ex.2: *The death penalty should be abandoned everywhere*

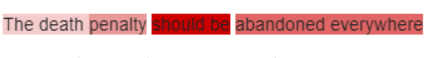


Figure 6: Sentence from MT

Ex.3: *My child went to a private all boys school and it was one of the best thing I ever did!*

[1] http://mpqa.cs.pitt.edu/lexicons/arg$_{l}exicon/$

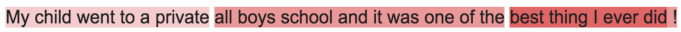


Figure 7: Sentence from WD

## 6. Results

We ran all the in-domain experiments and cross-domain experiments using our 3 models.
A closer analysis showed that except for the WTP and OC datasets which were the most noisy data sets in our corpora, our **BiLSTM+Attention+Arglex** model outperformed the other two models.

Table 2. **In domain Claim-F1 scores** for our three models compared with Daxenberger et al. The highlighted cells represent the best F1 score for each dataset

| Claim F1 Scores | Daxenberger et al. | BiLSTM + Dropout | BiLSTM + Attention | BiLSTM + Attention + ArgLex |
|---|---|---|---|---|
| MT | 41.8 | 55.4 | 50.2 | **64.4** |
| PE | 62.0 | 61.2 | 61.7 | **63.0** |
| WD | 24.5 | 23.9 | 24.8 | **26.23** |
| VG | 37.7 | 35.6 | 38.1 | **38.5** |
| OC | 22.4 | **25.2** | 25.0 | 24.9 |
| WTP | 28.5 | 28.5 | **28.9** | 28.4 |

We report weighted F1 score mainly because of the heavy imbalance in our dataset and in our test dataset specifically.

Table 3. **Weighted F1 scores** for our three models. Highlighted cells represent the best result for each dataset

| Weighted F1 Scores | BiLSTM. | BiLSTM + Attention | BiLSTM + Attention + ArgLex |
|---|---|---|---|
| MT | 79 | 81.6 | **83.0** |
| PE | 64.5 | 70.9 | **72.0** |
| WD | 46.5 | 46.43 | **48.6** |
| VG | 51.2 | 53.76 | **54.12** |
| OC | 75.0 | **81.0** | 74.6 |
| WTP | 69.0 | **74.5** | 72.3 |

## 7. Error Analysis

We present a thorough error analysis in this section to give a brief idea on how our models perform.

Below are sentences from the MT and WTP datasets which were **claims** but predicted as **non-claims** with our **BiLSTM** model

Table 4. **Cross-domain experiments**. Underlined numbers across diagonal are for the in-domain results and highlighted number in blue represent the best result for each dataset

| Claim F1 Scores | MT | PE | WTP | VG | OC | WD |
|---|---|---|---|---|---|---|
| MT | **64.4** | 13.4 | 5.9 | 17.6 | 9.25 | 13.16 |
| PE | 56.6 | 63.0 | 16.6 | 28.8 | 9.0 | **13.6** |
| WTP | 27.2 | 18.5 | 28.9 | 20.2 | 15.6 | 11.6 |
| VG | 51.2 | 33.7 | 22.4 | **38.5** | 18.6 | 11.37 |
| OC | 38.6 | 44.5 | **30.6** | 35.2 | 25.2 | 13.3 |

*The implementation of retirement at 63 is no longer socially sustainable.*

*Like I indicated, there is valid criticism, but it's probably has more to do with appearances than actual responsibility.*

On the contrary below are the examples which were correctly predicted as **claims**

*The EU should exert influence on the political events in Ukraine.*

*Your comment here is clearly biased and anything that is coming from a biased standpoint on either side of the issue should not be included.*

With our **BiLSTM + attention** model, our F1 scores improved and so did our predictions. This is evident from the fact that both the sentences from the MT and WTP datasets which were **claims** but mis-classified as **non-claims** with our **BiLSTM** model were classified correctly now.

*The implementation of retirement at 63 is no longer socially sustainable.*

*Like I indicated, there is valid criticism, but it's probably has more to do with appearances than actual responsibility.*

Below are two sentences from the MT and WTP corpora again. The **BiLSTM+attention** model couldn't identify these correctly but **BiLSTM+Attention+Arglex** model identified them correctly.

*Opening on Sundays and holidays would **therefore** help both customers and shops.*

*It may be **necessary** to revisit a semi protect then.*

Here 'therefore' and 'necessary' are elements from the arguing lexicon. Including these features improved our network predictions.

## 8. Github code repository

`https://github.com/tuhinjubcse/COMS4995`

This repository holds the code for creating 10 cross validations of the datasets as explained in the methods section, Python code for all three models and code for obtaining the weighted and claim F1-scores and the 6 data.txt files. We thank [3] for making their github repository `https://github.com/UKPLab/emnlp2017-claim-identification` public which helped us in this project .

## 9. Conclusion

We have successfully implemented word level attention on the BiLSTM deep learning architecture (initialized with random word embeddings) and hybrid deep learning model augmented with features from argumentation lexicon. We obtained results for in-domain and cross-domain experiments. Incrementally, we obtained better numbers after having applied attention. Further improvement was observed with our hybrid approach to this task of claim identification. Overall, we observed improved performance numbers in comparison to the performance reported in [3].

## 10. Future Work

In the future, we will work on Multi Task learning to support the hypothesis that weights trained on one dataset, shared with another can improve overall classification quality. We hope to incorporate Frame Net embeddings taking motivation from [2]. We also plan to add more linguistically inspired features as claim identification is a hard problem for deep learning models to learn. We want to incorporate context aware embeddings to improve a claim's support. We also plan to explore multitask learning for improving cross-domain claim identification by taking motivation from the work in [12]

## 11. Acknowledgements

The authors of the report would like to thank Professor Iddo Drori for his constructive comments and his help in shaping this project report.

## References

[1] K. Al-Khatib, H. Wachsmuth, M. Hagen, J. Köhler, and B. Stein. Cross-domain mining of argumentative text through distant supervision. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1395–1404, 2016.

[2] T. Botschen, H. M. Sergieh, and I. Gurevych. Prediction of frame-to-frame relations in the framenet hierarchy with frame embeddings. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 146–156, 2017.

[3] J. Daxenberger, S. Eger, I. Habernal, C. Stab, and I. Gurevych. What is the essence of a claim? cross-domain claim identification. *arXiv preprint arXiv:1704.07203*, 2017.

[4] S. Eger, J. Daxenberger, and I. Gurevych. Neural end-to-end learning for computational argumentation mining. *arXiv preprint arXiv:1704.06104*, 2017.

[5] D. L. Fan Zhang, Rebecca Hwa and H. B. Hashemi. Argrewrite: A web-based revision assistant for argumentative writings. In *In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations ,San Diego, CA, USA.*, pages 37–41, 2016.

[6] R. Levy, Y. Bilu, D. Hershcovich, E. Aharoni, and N. Slonim. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500, 2014.

[7] X. Ma and E. Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*, 2016.

[8] R. M. Palau and M.-F. Moens. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107. ACM, 2009.

[9] A. Peldszus and M. Stede. Joint prediction in mst-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948, 2015.

[10] R. Rinott, L. Dankin, C. A. Perez, M. M. Khapra, E. Aharoni, and N. Slonim. Show me your evidence-an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 440–450, 2015.

[11] S. Rosenthal and K. McKeown. Detecting opinionated claims in online discussions. In *Semantic Computing (ICSC), 2012 IEEE Sixth International Conference on*, pages 30–37. IEEE, 2012.

[12] A. Søgaard and Y. Goldberg. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 231–235, 2016.

[13] S. Somasundaran, B. Riordan, B. Gyawali, and S.-Y. Yoon. Evaluating argumentative and narrative essays using graphs. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1568–1578, 2016.

[14] C. Stab and I. Gurevych. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, 2014.

[15] J. R. Swapna Somasundaran and J. Wiebe. Detecting arguing and sentiment in meetings. In *SIGdial Workshop on Discourse and Dialogue*, pages 146–156, 2007.

[16] trudy govier. a practical study of argument, 7th edition. In *wadsworth, cengage learning*, 2010.

[17] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, 2016.