# Classifying Argument Component using Deep Learning on English Dataset

William Gunawan[1*] and Derwin Suhartono[2]
[1] Computer Science Department, BINUS Graduate Program, Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia.
[2] Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia.
E-mail: william.gunawan001@binus.ac.id, dsuhartono@binus.edu

**Keywords:** argumentation mining, argument component, bidirectional encoder representations from transformers

**Received:** January 14, 2023

*The study focuses on the argument component in argumentation mining, specifically examining claim and premise types. Various datasets exist for argumentation components, each with different classes. The study evaluates the performance of deep learning architectures, particularly using contextual embedding as the initial layer. Six datasets with diverse argument components are used for validation. The research provides a comprehensive comparison of deep learning architectures, combining multiple layers such as BERT or word embedding with LSTM, GRU, or CNN. The results and their implications are discussed in the concluding section of the journal. The study demonstrates significant results with the BERT-BiGRU-CRF architecture after conducting several experiments.*

*Povzetek: Študija preučuje komponente argumentacije pri rudarjenju argumentov, pri čemer je osredotočena na arhitekture globokega učenja in kontekstno vstavljanje.*

## 1 Introduction

Argumentation is a major element of human intelligence. The ability to argue is fundamental for humans to understand new problems, perform scientific reasoning, express, clarify and defend an opinion in everyday life [1]. Therefore, argumentative sentences frequently appeared in public spaces, namely social media debates, reviews, and scientific articles. Nevertheless, determining the meaning of an argumentative sentence requires complex processes and the use of deep learning could accelerate the completion of the task.

Natural language processing (NLP) techniques such as Argumentation mining could identify and classify the component of arguments contained in a writing. By focusing on the automatic identification of argument structures in natural languages [2], it has the capability to understand an argumentation structure so that the reasons for the opinions issued can be known [3]. In addition, the technique is not limited to understanding the meaning of each word. The advantage yields valid argumentation sentences because the arguments are supported by relevant facts [4]. Furthermore, a comprehension of the relationship between argumentative sentences is needed to get the meaning of the sentence [5].

Table 1: Each dataset with its argument component followed by the count of argument components from each class.

| Dataset | Type of Argument Component |
|---|---|
| Web discourse | Backing (205), Claim (183), Premise (499), Rebuttal (65), Refutation (23) |
| Persuasive essays | Claim (1,160), Major Claim (465), Premise (3,336) |
| Hotel reviews | Background (157), Claim (936), Implicit Premise (112), Major Claim (259), Premise (385), Recommendation (118) |
| News comments | Premise (4,294) |
| Various (Araucaria) | Premise (1,229), Claim (496) |
| Wiki discussions | Premise (1,299), Claim (1,039) |

Datasets on argumentative sentences are developed from time to time, as can be seen in Table 1. There are several components of the argument that are familiar. the components of the argument such as backing, rebuttal, and refutation which have been described in other studies. [6]. The majority of datasets consist of promises and claims arguments, meanwhile, Web and Hotel datasets contain additional arguments, namely Backing, Rebuttal, Refutation, Background, Implicit Premise, Major Claim, and Recommendation.

Numerous research has been accomplished in argument component classification topics with various datasets and approaches. They produce valuable results, unfortunately, some limitations are still shown. Social media [7], news [8], essays or articles [9], and Wikipedia discussions [10] are the datasets that are being used. Moreover, the approaches have a great range of algorithm complexity, they are using Support Vector Machine (SVM) with max entropy [11,12], deep learning [4,13], probabilities modeling [14], and Transformer model using BERT model [15].

This paper focuses on classifying argument components with various types of datasets using several deep learning architectures, specifically the BERT-related models. With the different number of classes in each dataset, it is expected to provide more insight from the obtained result on each model and dataset. Some previous works remarkably inspired this research. Firstly, the research uses MTL and STL on six datasets and provides an understanding that the amount of data and the diversity of classes in the existing datasets cannot provide the same improvement [2].

Secondly, the work with the approach of the use of contextual language models provides promising results in classifying argumentation components [16]. Thirdly, the approach of using BLSTM-CNNs model that handles sequence labeling data [17], and lastly, the research that uses the combination of BERT and Bidirectional RNN (Recurrent Neural Network) architecture [18- 20]. As a disclaimer, this study does not provide a comparison between the conducted research with the previous ones.

## 2 Related works

Conducted studies on argumentation mining, especially in the argument component, to come up with insightful ideas. Moreover, they apply diversified architecture and a great number of them provide tremendously. The research of argumentation mining started with the roots of philosophy [21]. The evolution of the Artificial Intelligence (AI) algorithm followed by the advantage technique in Machine Learning (ML) produces impressive progress that attracts the scientific community [22]. The use of Machine Learning techniques combined with statistical knowledge such as maximum entropy and the rules of Context Free Grammar (CFG) obtained a promised result [11].

Plentiful approaches are used by researchers using the combination of NLP discipline. Great improvement in the argumentation mining area using semantic textual similarity (STS) combined with textual entailment [23] and the research uses the combination of 8 features containing structural, lexical, syntactic, contextual, indicator, embedding, probability, and similarity [24]. It shows that the interaction in arguments can be used to recognize the argument. Another approach is carried out by identifying the component of the argument using multiclass classification with a Support Vector Machine (SVM) followed by identifying the structure of the argument [25]. And the research that applied Word2Vec and semi-supervised learning to the argument data with sequence structure in Greek [8].

The advantage technique of Deep Learning (DL) has been shown and enlarged the research area in argumentation mining, especially on the component of the argument. By comparing the usage of Bidirectional Long Short-Term Memory using Single-Tasking Learning (STL) and Multi-Tasking Learning (MTL). Experiment on six argumentation mining datasets with the same model. It shows that the complex model can beat a shallow one, which comes from the results that MTL overcomes STL on every dataset [2]. Besides that, the imbalanced data of the component of the argumentation problem has become one of the research branches of argumentation mining. Using SVM and Partial Tree Kernel (PTK), show that imbalanced data can be solved [26].

Mainly, Argumentation mining is like the other NLP tasks. Argumentation mining has been studied using Transformer models [15,16]. Not limited to classifying argumentation components, the Transformer model is also used to classify the relations of the argument [13], and even the use of the Transformer model to summarize the argument gives a promising result [27].

Table 2: Summary of related works results.

| Ref. No. | Year | Data | Technique | Results |
|---|---|---|---|---|
| [2] | 2018 | Table 1 | MTL and STL using BLSTM | Table 3. Column 2 |
| [13] | 2020 | Corpus US2016 & Moral Maze cross-domain | Transformer model | 70% for US2016 and 61% for Moral Maze cross-domain |
| [15] | 2020 | Extended MEDLINE Corpus | Fine-tuning SciBERT | F1-score 87% |
| [25] | 2014 | Persuasive Essays | Multi-class SVM with features selections | F1-score 72.2% |
| [26] | 2019 | IBM Topic Corpus | SVM and partial tree kernel (PTK) | F1-score 74% |
| [27] | 2021 | IBM Debater(R)-ArgKP | Text-to-Text Transfer Transformer | F1-score 98.5% |

Table 2 provides a concise summary based on similar research that has been published. Most of the research that has been done has only focused on one dataset. This is what underlies the researchers to conduct this research. Following the given knowledge from previous research, we developed several Transformer-based models, several deep learning models which represent the sequence-to-sequence model, and a combination of BERT and deep

learning models. This study will provide an overview of how each Transformer-based model is able to study the six datasets that have different labels and uneven distribution of data. The comparison between models will be given in section 4.

# 3 Proposed method

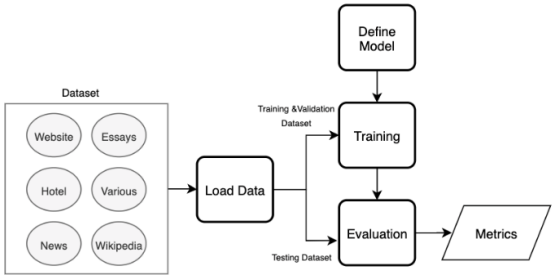


Figure 1: Research Frameworks.

With several conducted experiments, the proposed model is constructed of several deep learning models and uses frameworks as shown in Figure 1. Initially, the datasets are loaded, then they are trained with a predefined model, and finally compared with the testing metrics.

## 3.1 Dataset

The experiments utilize six different datasets that have been preprocessed. They are transformed into a proposed token level using BIO tags [28]. The BIO itself stands for Begin, Inside follows the begin tag, and Outside for the words that are not in the classes. Furthermore, the amount of data is not distributed well and each dataset has different labels. Hence, the data is trained and evaluated separately.

Table 3: The dataset used in this research with the total document is used for training, validation, and testing.

| Dataset | Total Training Data | Total Validation Data | Total Testing Data | Total Training Token |
|---|---|---|---|---|
| Web | 136 | 60 | 338 | 21,542 |
| Essays | 108 | 45 | 598 | 21,013 |
| Hotel | 138 | 64 | 36 | 21,042 |
| News | 196 | 86 | 1,645 | 21,031 |
| Various | 192 | 81 | 263 | 21,084 |
| Wikipedia | 130 | 57 | 954 | 21,066 |

The table above presents the sources of the original datasets that are expected to have a great result in identifying the argumentation component. The dataset consists of Various (Araucaria) [29], Wikipedia Discussions [10], Hotel Reviews [30], Web discourse [31], News Comments [32], and Persuasive Essays [25]. Besides, the total training data, total validation data, and total testing data refer to the number of documents. Moreover, our research uses 21K training data that has already been used in previous research [2].

Table 4: Label Distribution.

| Dataset | Label |
|---|---|
| Web | Backing (2.557), Claim (953), Premise (5.733), Rebuttal (529), Refutation (472), Other (11.298) |
| Essays | Claim (3.387), Major Claim (1454), Premise (9.539), Other (6.633) |
| Hotel | Background (1.495), Claim (8.110), Implicit Premise (1.626), Major Claim (1.402), Premise (4.574), Recommendation (1.241), Other (2.594) |
| News | Premise (10.999), Other (10.032) |
| Various | Premise (9.817), Claim (3.355), Other (7.912) |
| Wikipedia | Premise (5.340), Claim (1.706), Other (14.020) |

The number of each label in each dataset can be seen in the table above. Each dataset has a relatively broad distribution of labels, but the Web and Wikipedia datasets stand out as having significantly fewer positive than negative labels. The positive labels and the negative labels are almost evenly distributed in the other datasets.

## 3.2 Deep learning architectures

Several deep-learning approaches are applied in this research. Furthermore, the models that solve sequence problems are preferred to the traditional machine learning algorithm. The architecture models implement three different types of embedding layers. The first layer is the BiGRU model without pre-trained word embedding. The second ones are BiGRU and BLSTM-CNNs models that use Glove with 200 dimensions with six billion tokens as pre-trained word embedding. Lastly, the models with contextual embeddings, such as BERT, DistilBERT [33], and BERT-BiGRU-CRF.

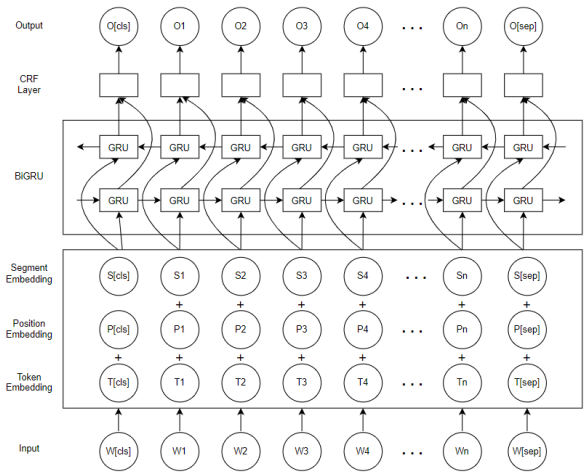


Figure 2: BERT-BiGRU-CRF Architecture.

On the predicted layer, BERT-BiGRU-CRF uses CRF, while the other models do not use contextual embedding with dens layer and Softmax. Uncased pre-trained models are implemented on the model that includes contextual embedding and every available token has been lowercase. Five CNN layers are set in parallel to the word embedding layer on the BLSTM-CNNs architecture model. Later the CNN layers will be concatenated and continued with two BLSTM layers of 200 units each.

For BiGRU models that either do not apply pre-trained word embedding, they use two BiGRU layers with 200 units each. On the BERT-BiGRU-CRF model, 2 BiGRU layers with 200 units each and the CRF layer as the prediction layer are set sequentially. In general, every

architecture model uses a dropout layer with a value of 0.5 after the embedding layer.

# 4 Experiment

Two treatments are equally applied to the proposed architecture models. First, the model utilizes six datasets and is accomplished in two batch sizes, namely 8 and 12 that yield 72 experiments in total. Second, the model uses a 512-sequence length which determines the token value, if it is less than 512 it will be filled with padding otherwise the rest value will be ignored. Besides that, other procedures are uniformly applied for each experiment.

For instance, Adam optimizers are implemented with a learning rate starting from 3x10-4 with epsilon value 10-8 and the data will be trained using epochs as many as 100. Early stopping, one of the regularizations is chosen to the outcome of the overfitting. It is configured with four maximum errors, meaning that the training will be stopped if the iteration is unable to reduce the loss value four times.

# 5 Results and discussion

Table 5: Macro-F1 for AM component.

| Dataset | Model | | | | | | |
|---|---|---|---|---|---|---|---|
| | Previous Research [2] | BiGRU | BiGRU + Gloves Embedding | BLSTM-CNNs + Glove Embedding | BERT | DistilBERT | BERT-BiGRU-CRF |
| Web | 0.234 | 0.173 | 0.265 | 0.262 | 0.066 | 0.075 | 0.299 |
| Essays | 0.605 | 0.257 | 0.612 | 0.594 | 0.604 | 0.504 | 0.655 |
| Hotel | 0.479 | 0.214 | 0.435 | 0.458 | 0.476 | 0.427 | 0.504 |
| News | 0.577 | 0.477 | 0.632 | 0.619 | 0.558 | 0.407 | 0.677 |
| Var | 0.474 | 0.356 | 0.505 | 0.354 | 0.481 | 0.368 | 0.59 |
| Wiki | 0.325 | 0.305 | 0.38 | 0.397 | 0.312 | 0.251 | 0.434 |

This research does not have any comparison with other research, hence, the result using a full dataset will become divergent. All the models will be evaluated using the Macro-F1 score because we focused on a positive class. Batch size 8 experiments produce better results than the 12 ones for many of the datasets except for the Web dataset. Web dataset is difficult to learn and provides unacceptable results on every model. In contrast, News, and Essays datasets are easier to understand. Table 4 shows the comparison results on each dataset.

Following the results from Table 5, can assume that BERT-BiGRU-CRF has overcome other models include with the previous research with the same dataset distribution. But the results are not significant to some datasets. Poor results are obtained on datasets that have many classes. However, the dataset with 2 components such as Various (araucaria) and Wikipedia increased by about 10% more than the other experiment. The imbalanced data has a big role in the results. Mostly, each dataset that has an imbalanced argument component gives a bad result to the average or micro F1-Score. In this case, the class imbalance that occurs is the uneven distribution between the positive class and the negative class and between each positive class.

The model runs in a very small iteration because of the small amount of data, the model learns about 14 to 24 variations based on the total training data over the given batch size. But the BERT model reaches the 50 epochs without being penalized by the early stopping function. From Table 4, it can be concluded that word embedding is very influential. Moreover, in this case, BERT based model used using vanilla BERT does not give a good result for Hotel and Web dataset that has various classes. The combination between BERT and GRU extended by CRF can give a better result for every dataset.

In general, the authors discovered a number of flaws in the native Transformer model, including the model's inability to perform well on data with uneven distribution, like the Web and Wikipedia. With other datasets, however, it can accommodate deep learning hybrid models built on Transformers and non-BERT models. The hybrid approach also can't perform well on data from the Web and Wikipedia.

As a result, the pattern of subpar findings for the two data is based on a comparison of the relatively large number of positive and negative labels. To further support the analysis, it should be noted that the Hotel dataset, despite a small distribution of positive label data, still yields promising results due to the distribution between

# 6 Conclusion

In short, the experiment result shows that the BERT model achieves better performance on the data with a small number of classes, yet poor on the data with numerous classes such as Hotel and Web datasets. Besides, BLSTM-CNNs model obtains stable performance, nevertheless insignificant.

Several things can be concluded after doing this research, namely:

- BERT-BiGRU-CRF overcomes all the experiments for each dataset.
- BERT and DistilBERT models deliver poor results on the diverse imbalance classes of datasets. However, they produce the same result as BLSTM-CNNs or BiGRU with Glove embedding on a small number of class datasets.
- CNN in BLSTM-CNNs model results steadily. It is proven by the comparison of BiGRU with Glove embedding.
- Imbalance class of argument component results worse on the small amount label than the large one. For example, the refutation argument in the Web dataset only consists of 5% compared to the claim argument component.

Several improvements in features engineering and parameter tuning could potentially be advancing the research. The first is to increase the experiment with another BERT model such as Big Bird [34]. The second is to use the hybrid model of BERT followed by BLSTM-CNNs in one architecture. The last is to use bigger word embeddings for the non-BERT model to compare this research.

# 7 References

[1] Mochales, R. & Moens, M. F. (2011). *Argumentation Mining. Artificial Intelligence and Law.* 19. 1-22. doi:10.1007/s10506-010-9104-x.

[2] Schulz, C., Eger, S., Daxenberger, J., Kahse, T., & Gurevych, I. (2018). *Multi-Task Learning for Argumentation Mining in Low-Resource Settings.* NAACL- HLT. doi:10.18653/v1/n18-2006.

[3] Lawrence, J. & Reed, C. (2019). *Argument Mining: A Survey. Computational Linguistics.* 45. 765-818. doi:10.1162/COLI_a_00364.

[4] Suhartono, D., Gema, A. P., Winton, S., David, T., Fanany, M. I., Arymurthy, A. M. (2020). *Argument annotation and analysis using deep learning with attention mechanism in Bahasa Indonesia.* Journal of Big Data. 7(90). Springer. doi:10.1186/s40537-020-00364-z.

[5] Gema, Aryo & Winton, Suhendro & David, Theodorus & Suhartono, Derwin & Shodiq, Muhsin & Gazali, Wikaria. (2017). *It Takes Two To Tango: Modification of Siamese Long Short Term Memory Network with Attention Mechanism in Recognizing Argumentative Relations in Persuasive Essay*. Procedia Computer Science. 116. 449-459. doi:10.1016/j.procs.2017.10.036.

[6] Hunter, Anthony. (2007). *Elements of Argumentation.* 4. 10.1007/978-3-540-75256-1_3.

[7] Lippi, M., & Torroni, P. (2016). *Argumentation Mining: State of the Art and Emerging Trends.* ACM Trans. Internet Techn., 16, 10:1-10:25. doi:10.1145/2850417.

[8] Sardianos, C., Katakis, I.M., Petasis, G., & Karkaletsis, V. (2015). *Argument Extraction from News.* ArgMining@HLT-NAACL. doi:10.3115/v1/w15-0508.

[9] Stab, C., Kirschner, C., Eckle-Kohler, J., & Gurevych, I. (2014). *Argumentation Mining in Persuasive Essays and Scientific Articles from the Discourse Structure Perspective.* ArgNLP.

[10] Biran, O., & Rambow, O. (2011). *Identifying Justifications in Written Dialogs by Classifying Text as Argumentative.* Int. J. Semantic Comput., 5, 363-381. doi:10.1142/s1793351x11001328.

[11] Palau, R., & Moens, M. (2009). *Argumentation mining: the detection, classification and structure of arguments in text.* ICAIL. doi:10.1145/1568234.1568246.

[12] Lippi, M., & Torroni, P. (2015). *Argument Mining: A Machine Learning Perspective.* TAFA.

[13] Ruiz-Dolz, R., Barberá, S.H., Alemany, J., & García-Fornes, A. (2020). *Transformer- Based Models for Automatic Identification of Argument Relations: A Cross- Domain Evaluation*. ArXiv, abs/2011.13187. doi:10.1109/mis.2021.3073993.

[14] Culotta, A., McCallum, A., & Betz, J. (2006). *Integrating Probabilistic Extraction Models and Data Mining to Discover Relations and Patterns in Text.* HLT-NAACL. doi:10.3115/1220835.1220873.

[15] Mayer, T., Cabrio, E., & Villata, S. (2020). *Transformer-Based Argument Mining for Healthcare Applications.* ECAI.

[16] Hidayaturrahman, Dave, E., Suhartono, D., & Arymurthy, A. M. (2021). *Enhancing argumentation component classification using contextual language model.* Journal of Big Data, 8(1), [103]. doi:10.1186/s40537-021-00490-2.

[17] Gunawan, W., Suhartono, D., Purnomo, F., & Ongko, A. (2018). *Named-Entity Recognition for*

*Indonesian Language using Bidirectional LSTM-CNNs.* Procedia Computer Science, 135, 425-432. doi:10.1016/j.procs.2018.08.193.

[18] Yu, Q., Wang, Z., & Jiang, K. (2020). *Research on Text Classification Based on BERT-BiGRU Model.* Journal of Physics Conference Series, 2021, vol. 1746, no. 1. doi:10.1088/1742-6596/1746/1/012019.

[19] Tobias, M., Elena, C., & Serena, V. (2020) *Transformer-based Argument Mining for Healthcare Applications.* In Proceedings of the 24th European Conference on Artificial Intelligence (ECAI 2020), Santiago de Compostela, Spain. doi:10.3233/faia325.

[20] Qin, Q., Zhao, S., & Liu, C. (2021). *A BERT-BiGRU-CRF Model for Entity Recognition of Chinese Electronic Medical Records.* Complexity. 2021. 1-11. doi:10.1155/2021/6631837.

[21] Toulmin, S. (2008). *The Uses of Argument, Updated Edition.*

[22] Lytos, A., Lagkas, T., Sarigiannidis, P., & Bontcheva, K. (2019). *The evolution of argumentation mining: From models to social media and emerging tools.* ArXiv, abs/1907.02258. doi: 10.1016/j.ipm.2019.102055.

[23] Boltuzic, F., & Šnajder, J. (2015). *Identifying Prominent Arguments in Online Debates Using Semantic Textual Similarity.* ArgMining@HLT-NAACL. doi:10.3115/v1/w15-0514.

[24] Winata, R., Haryono, E. G., Suhartono, D. (2021). *Toward Better Argument Component Classification in English Essays.* ICIC Express Letters. vol. 12. 111-119.

[25] Stab, C., & Gurevych, I. (2014). *Identifying Argumentative Discourse Structures in Persuasive Essays.* EMNLP. doi:10.3115/v1/d14-1006.

[26] Kusmantini, H.A., Asror, I., & Bijaksana, M.A. (2019). *Argumentation mining: classifying argumentation components with Partial Tree Kernel and Support Vector Machine for constituent trees on imbalanced persuasive essay. doi:10.1088/1742-6596/1192/1/012009*

[27] Harly, W., Kwee, R. H., Suhartono, D. (2022). Quantitative Argument Summarization Using Text-to-Text Transfer Transformer. ICIC Express Letters Part B: Applications. vol. 13, 749-756. doi:10.24507/icicelb.13.07.749.

[28] Ratinov, L., & Roth, D. (2009). *Design Challenges and Misconceptions in Named Entity Recognition.* Conference on Computational Natural Language Learning. doi:10.3115/1596374.1596399.

[29] Reed, C., Palau, R., Rowe, G., & Moens, M. (2008). *Language Resources for Studying Argument.* LREC.

[30] Liu, H., Gao, Y., Lv, P., Li, M., Geng, S., Li, M., & Wang, H. (2017). *Using Argument-based Features to Predict and Analyse Review Helpfulness.* ArXiv, abs/1707.07279. doi:10.18653/v1/d17-1142.

[31] Habernal, I., & Gurevych, I. (2017). *Argumentation Mining in User-Generated Web Discourse.* Computational Linguistics, 43, 125-179.

[32] Habernal, I., Wachsmuth, H., Gurevych, I., & Stein, B. (2017). *The Argument Reasoning Comprehension Task.* ArXiv, abs/1708.01425. doI:10.1162/coli_a_00276.

[33] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.* ArXiv, abs/1910.01108.

[34] Zaheer, M., Guruganesh, G., Dubey, K.A., Ainslie, J., Alberti, C., Ontañón, S., Pham, P., Ravula, A., Wang, Q., Yang, L., & Ahmed, A. (2020). *Big Bird: Transformers for Longer Sequences.* ArXiv, abs/2007.14062.