An attempt to combine features in classifying argument components in persuasive essays

Yunda Desilia, Velizya Thasya Utami, Cecilia Arta, Derwin Suhartono
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
{yunda.desilia, velizya.utami, cecilia.arta}@binus.ac.id, dsuhartono@binus.edu

ABSTRACT
So far, several approaches have been done in detecting and classifying argumentation in persuasive essays. In this paper, we proposed some new features on top of the state-of-the-art researches in argumentation mining. We grouped 68 features into 8 categories; they are structural, lexical, indicators, contextual, syntactic, prompt similarity, word embedding, and discourse features. Instead of handcrafted features, we utilized word embedding as the feature. At the end of this paper, we presented the comparison between each group of features to classify the argument components. 402 persuasive essays were utilized. We found that structural features were the most significant feature while discourse features were not. After combining all features, we obtained 79.96% as the accuracy; it was slightly outperforming the state-of-the-art accuracy which was 77.3%.

Keywords
argument component, feature, word embedding, argumentation mining, persuasive essay

1. INTRODUCTION
Argumentation is a process of building arguments, exchanging arguments, and evaluating arguments in terms of interaction with the other arguments. An argument is a set of premises or evidence/fact, which are given to support the claim (Palau and Moens, 2009). The objective of argumentation is to make the audiences believe the idea, thought, or opinion stated are true and proved. Argumentation mining aims to detect the arguments in a text document, relation between them, and internal structure of each argument. By integrating argumentation mining in writing environments, human will be able to inspect their text for plausibility and to improve the quality of their argumentation.

A minimum definition of an argument is a set of statement that consists of 3 parts: conclusion, premises, and inference (Walton, 2009). On the other hand, it is stated that argument is a statement with 3 components: claim/point of view that is argued, actual argument/evidence, a statement that links first claim to the argument and makes sure the function of argument can be understood. (Moens, 2014)

Palau (2008) stated that argumentation detection can help to facilitate understanding of argumentation paragraph, demonstrate a good identification for important information, increase the possibility of indexing implementation or document searching, represent reasoning system. The classification of argument component and visualization has several advantages, such as to show clear, strong, and structured/organized arguments. Besides, it also facilitates evaluation of opinion, facilitates understanding of other's opinions, helps in giving the teaching of general thoughts, and helps in teaching critical thinking. Thus, having a better accuracy in classifying argument components becomes a compulsory problem.

In this work, we proposed some new features on top of the state-of-the-art research in argumentation mining. We implemented 68 sub-features that are grouped into 8 main categories of features. They are structural, lexical, contextual, indicators, prompt similarity, syntactic, word embedding, and discourse. We also provided accuracy comparison to previous systems that were related to our work. We propose approach that consists of two main steps in our research. First, we did component identification, which include a process of identification and detection of argument component. We separated argumentative text units from non-argumentative text units and also identified the presence of argument component. Secondly, we did component classification, which include classification process of argument component type into major claim, claim, premise, or non-argumentative.

2. RELATED WORKS
There are several works that are related with this research, specifically in the field of argument detection and classification. Moens, Boiy, Palau, and Reed (2007) did a research of automatic detection of arguments in legal texts. They used lexical, syntactic, semantic, and discourse features. In this research, they used Araucaria corpus as the dataset and Multinomial naïve Bayes and maximum entropy model as the classifiers. As the result, they obtained 74% accuracy of all features extraction in the variant of texts and 68% in legal texts. The detection and classification of argument component and the identification of argument structure was proposed by Palau and Moens (2009). They used Araucaria corpus and European Court of Human Rights (ECHR) as the data and feature extraction as the method. This research obtained 73% of accuracy in Araucaria and 80% of accuracy in ECHR. On the other hand, the accuracy was 74.07% for premise and conclusion classification and it yielded 60% for detecting the argument structure. Lippi and Torroni (2015) proposed several methods to detect claims. They used IBM corpus dataset and 90 persuasive essays. As the result, they achieved 71.4% of accuracy in the 90 persuasive essays and 20.6% in IBM corpus. Al-Khatib et al. (2016) proposed a distant supervision approach in classifying argumentative parts in text automatically from online debate portal. They used corpus of Webis-debate-16 and did a cross-domain comparison with 90 persuasive essays and web discourse corpus. This research achieved 66.8% of accuracy in 90 persuasive essays corpus, 87.7% of accuracy in web discourse corpus, and 91.8% of accuracy in Webis-debate-16. For the experiment of cross-domain comparison, the highest accuracy was obtained by web discourse corpus tested in Webis-debate-16, which reached 84.4% of accuracy.

The other focus to classify the arguments by identifying argumentation schemes was done by Feng and Hirst (2011). They used Araucaria database, features extraction, and two methods of classification. The features used in this research were general and scheme-specific features. The highest accuracy was 90.8% in scheme target of reasoning while the lowest accuracy was 63.2% in scheme target of classification for one-against-others-classification. For pairwise classification, the highest accuracy was 98.3% in scheme target of classification-reasoning and the lowest accuracy was 64.2% in scheme target of classification-consequences.

To identify the argumentative discourse, some researchers did annotation study to create the corpus. Stab and Gurevych (2014a) did the annotation study and created corpus of 90 persuasive essays. They continued the research by identifying the argument component and the argumentative relations in persuasive essays. Support Vector Machine (SVM) was used and it obtained 77.3% of accuracy with structural feature as the best performing feature. On further research, they created an approach to parse the argumentation structures in persuasive essays (Stab and Gurevych, 2016). They created corpus of 402 persuasive essays and extracted the features to identify the argument component, classified the argument component, identified the argumentative relation, tree generation, and stance recognition. They obtained 77.3% of accuracy and structural was the best performing features. They also proposed approach to recognize the absence of opposing arguments in persuasive essays. They used both corpus of 90 persuasive essays and 402 persuasive essays. As the result, they got 75.6% of accuracy. The combination of unigrams, production rules, and adversative transitions obtained the highest accuracy among all of combinations. Habernal and Gurevych (2016) annotated and analyzed the arguments automatically in user-generated web discourse by extracting 5 (five) feature sets to detect the argument component. As the result, they obtained 75.4% of accuracy.

Some researchers focused on the approach to identify the argumentation structures. Peldszus (2014) proposed an approach to identify argumentation structures in micro text automatically with the various level of granularity. They used 115 micro text as the dataset and extracted the features and did a comparison with some types of classifiers. The most outperformed classifiers were Support Vector Machine (SVM) and Maximum Entropy Classifiers (MaxEnt). SVM obtained 64% of accuracy and MaxEnt obtained 63% of accuracy. The best features to obtain the high accuracy were lemma unigrams and lemma bigrams. Lawrence and Reed (2015) proposed 3 (three) methods to extract argumentation structures. They used AIFdb corpus and implemented discourse indicators, topic similarity, and schematic structure as the methods. The combination of those methods reached 83% of accuracy with the best performing feature was schematic feature.

Further implementation of argumentation detection and classification, such as accessing the quality of arguments have been done by some researchers. Wachsmuth, Al-Khatib, and Stein (2016) investigated mining structure to access the argumentation quality of persuasive essays. They used corpus that contains essays from International Corpus of Learner English, extracted the features, and classified the argument component into ADU types: thesis, conclusion, premise, and none. They obtained 74.5% of accuracy with the sentence position as the best performing feature.

3. METHODS

3.1 Data

We utilized a corpus of persuasive essays compiled by Stab and Gurevych (2016). It consists of 402 annotated persuasive essays with different kind of topics. This corpus contains argument component annotation in the clause-level as well as argumentative relations and argument structure in a different level of discourse. It also contains annotation about major claim, claim, premise in each of essay and consists of 7.116 sentences with 147.271 tokens.

3.2 Current Features

We implemented 68 sub-features that were categorized into 8 groups: structural, lexical, indicators, contextual, syntactic, prompt similarity, word embedding, and discourse features. The features described in this section were combined from some researches in argument components classification.

3.2.1 Structural Features

Structural features are features that identified argument component based on structure of the text. Covering sentence is a sentence that contains the argument component in it. Structural includes 3 sub-features, which are token statistics, location, and punctuation. For token statistics, we defined the number of tokens from argument component, the number of tokens from covering sentence, the number of tokens preceding and following an argument component in the covering sentence, the token ratio between covering sentence and argument component, the number of tokens from covering paragraph, the number of covering sentences preceding and following paragraph, the token ratio between covering sentence and covering paragraph, the token ratio between covering sentence and essay, the average number of token at sentence, the ratio and a Boolean feature that indicates if the argument component covers all tokens of its covering sentence as token statistics features. For location, we defined a set of location-based features for exploiting the structural properties of essay. 4 Boolean features that indicate if the argument component is present in the introduction or conclusion of an essay and if it is present in the first or the last sentence of a paragraph. Secondly, we add the position of the covering sentence in the essay and the position of the covering sentence in the paragraph as a numeric feature. We also count the ratio of covering sentence and paragraph, the ratio of covering sentence and essay, and the ratio of paragraph and essay. For punctuation, we define a set of punctuation-based feature to identify characteristics of argument component. This features will return the number of punctuation marks of the covering sentence and the number of punctuation marks of the argument component, the number of punctuation marks preceding and following an argument component in its covering sentence and a Boolean feature that indicates whether the sentence is closed with a question mark or not.

3.2.2 Lexical Features

These features are defined by N-grams, POS N-grams, verbs, adverbs, modals auxiliary, comparative and superlative adjective, the ratio of pronouns, and word couples.

3.2.3 Indicator Features

Boolean features indicating the presence of question indicators, time indicators, evidence indicator, conclusion indicator, compare-and-contrast, and cue phrases. We used 55 discourse markers as well and modelled each as a Boolean feature set to true if one of them is present in the covering sentence. The discourse markers were taken from the Penn Discourse Treebank 2.0 Annotation Manual (Prasad et. al., 2007). Furthermore, we also define 4 (four)

Boolean features that indicate the presence of type indicators including forward indicators, backward indicators, thesis indicators and rebuttal indicators. In addition, we defined 5 (five) Boolean features to identify possessive pronoun (I, me, mine, myself, my) in covering sentence.

3.2.4 Contextual Features

These features return the number of punctuations, number of tokens and sub-clauses from the sentence preceding and following the covering sentence, the number of covering sentence preceding and following the covering sentence. We also defined Boolean features indicating the presence of modal verbs, question indicator, comparative and superlative adjective, and type of indicators. In addition, we defined 4 (four) Boolean features and numeric that indicate if the shared noun and shared verb is present in the introduction or conclusion of an essay.

3.2.5 Syntactic Features

We count the number of sub-clauses in each sentence and return numeric value. We also count the depth of parse tree, extract the production rules, and identify whether the sentence is in past tense, present tense, or not in both.

3.2.6 Prompt Similarity Features

These features were created to count the similarity of cosine value between current sentence and the prompt, with the first sentence in each paragraph, with the last sentence in each paragraph, with its preceding sentence, and with its following sentence.

3.2.7 Word Embedding Features

They were created to count the vector representation of each word. Glove was used to obtain the vector representation for each word. We count the average of vector values per argument component.

3.2.8 Discourse Features

We implemented discourse doubles, which return: (1) count of explicit and implicit relation in a sentence and then return the count of which type present the most, (2) the ratio of explicit and implicit relation. Explicit discourse connectives are drawn primarily from well-defined syntactic classes, while implicit discourse connectives are inserted between paragraph-internal adjacent sentence pairs not related explicitly by any of the syntactically defined set of explicit connectives.

3.3 Additional Features

To explore further in classifying argument components, we defined some features which are quite promising to boost the accuracy of classification. Our additional features included 7 main features, which were structural, lexical, indicators, contextual, syntactic, prompt similarity, and discourse features.

- Structural features were number of token in covering paragraph, number of preceding and following covering sentence in covering paragraph, and position of covering sentence in paragraph.
- Lexical features were POS N-grams and word couples.
- Indicator features were forward, backward, rebuttal, thesis indicators, and cue phrases.
- Contextual features were type of indicators in context, number of shared noun and shared verb that are present in introduction and conclusion in essay, and 4 binary features that indicates shared noun and verbs that are present in introduction or conclusion in essay.
- Syntactic feature was POS distribution.
- Prompt similarity feature was the similarity of cosine value between current sentence with the prompt.
- Word embedding feature was defined to extract the vector representation of each word.

4. RESULTS AND DISCUSSION

4.1 Performance

There are 8 categories of features that were implemented for the features extraction: structural, indicator, contextual, lexical, syntactic, prompt similarity, word embedding, and discourse with total of 68 sub-features. We used Support Vector Machine (SVM) as classifier by using 10-folds cross validation and utilized a corpus of 402 annotated persuasive essays by Stab and Gurevych (2016).

The accuracy result of this system was 79.96%. It indicated that a higher accuracy was achieved in comparison to the argument component detection and classification systems conducted in the previous works as shown in Table 1. Even though this comparison did not show a proper objective evaluation due to task differences among them, our accuracy was quite promising to surpass previous works, especially to Stab and Gurevych (2014b).

Table 1. Previous works performance

Related Work	Accuracy
Palau and Moens (2007)	74%
Palau and Moens (2009)	74.04%
Stab and Gurevych (2014b)	77.3%
Lippi and Torroni (2015)	71.4%
Stab and Gurevych (2016)	77.3%
Wachsmuth, Al-Khatib, and Stein (2016)	74.5%
Habernal and Gurevych (2016)	75.4%
Al-Khatib et al. (2016)	66.8%

Table 2. Confusion matrix of the system accuracy results (SVM) for argument component classification

	MC	Cl	Pr	No
MC	**578**	130	43	0
Cl	226	**309**	970	1
Pr	28	147	**3656**	1
No	0	0	0	**1638**

Table 2 explains that the system correctly identifies 578 major claims (MC), 309 claims (Cl), 3656 premises (Pr), and 1638 non-argumentative (No). The errors occurred in identifying claims. Most of them were identified as premise. The accuracy in identifying each component was 76.96% for major claim, 20.52% for claim, 95.41% for premise, and 100% for non-argumentative. We guessed the accuracy to identify claims was very low due to class imbalance where claim had the lowest amount of data. Beside using 10-folds cross validation for training, we also conducted experiments using 5-folds cross validation with 79.74% accuracy.

Table 3. Previous works performance

Feature Name	Accuracy	Feature Name	Accuracy
Structural	77.83%	Syntactic	51.35%
Indicator	54.73%	Prompt Similarity	54.79%
Contextual	63.10%	Word Embedding	49.46%
Lexical	61.06%	Discourse	49.41%
All Features	**79.96%**		

We conducted experiments by using each feature group to capture which feature sets were significant in classifying the argument components. Based on Table 3, the best feature set to classify argument components is structural feature with 77.83% accuracy result. Contextual and lexical features consecutively were the next significant features among all.

4.2 Combining the Features

We attempted to combine all features as the next experiment. It was to identify which features combination has the best and the least impact in improving the system's accuracy.

Table 4. Accuracy result of combination of feature without one feature category in system

Feature Name	Accuracy	Feature Name	Accuracy
Without Structural	69.74%	Without Syntactic	78.21%
Without Lexical	77.72%	Without Prompt Similarity	79.93%
Without Indicators	77.98%	Without Word Embedding	78.48%
Without Contextual	78.05%	Without Discourse	79.98%

Based on Table 4, we can conclude that the most influential feature is structural, because all combination of features without structural has the lowest accuracy result with 69.74%, while the least influential feature is discourse as without discourse feature, the accuracy result is 79.98%.

From 8 trials of features combination, 7 of them showed significant accuracy, where 7 of them achieved an accuracy of 77.7% to 79.9%. This result indicates that the accuracy achieved by the combination of features produces higher accuracy compared to the accuracy of previous works (Table 1). In addition, we can see from the experiments that the accuracy of the system significantly decreased as a result of the feature extraction without structural features. Therefore, we also did an experiment with combination of 3 (three) features that achieved the highest accuracy, i.e. structural, lexical, and contextual features which produced 77.87% as the accuracy result.

4.3 Comparing Each Group of Features

We conducted other experiments by comparing system's accuracy among implementation by using the features presented by Stab and Gurevych (2014b), handcrafted features proposed by authors, and additional features from previous works. The system was trained using the same corpus consisting 402 annotated persuasive essays compiled by Stab and Gurevych (2016).

Stab and Gurevych (2014b) implemented structural, indicator, contextual, lexical, and syntactic features with total of 28 sub-features. Our system's accuracy result using features extraction based on Stab and Gurevych (2014b) is 76.32% (Table 5), while the original accuracy result of their research was 77.3% by using 90 persuasive essays where the highest accuracy is achieved by structural features. The result's difference can be caused by the different number of the training data.

Table 5. Accuracy result of implementation features by Stab and Gurevych (2014b)

Feature Name	Accuracy
Structural	74.33%
Indicator	61.11%
Contextual	52.38%
Lexical	58.69%
Syntactic	50.94%
All Features	**76.32%**

We proposed some handcrafted features to develop algorithm to identify and classify argument components and to increase the accuracy of system. This experiment used 24 sub-features which produced 68.46% as the accuracy result. In addition, we ran the system using each feature's category to identify each feature's performance (Table 6).

Table 6. Accuracy result of proposed handcrafted features

Feature Name	Accuracy
Structural	63.81%
Indicator	49.45%
Contextual	59.94%
Lexical	49.58%
Prompt Similarity	54.70%
Discourse	49.43%
All Features	**68.46%**

From the result presented in Table 6, the system achieved 68.46% accuracy with the highest accuracy achieved by structural features which followed by contextual and prompt similarity features as the second and the third most performing features.

The experiments also implemented additional features which were obtained from previous works conducted from state-of-the-art researches. There are 16 additional sub-features implemented in this scenario. Based on Table 7, the system achieved 71.08% of accuracy with the most significant accuracy was achieved by structural and contextual features. Word embedding feature was less performing feature in this experiment.

Table 7. Accuracy result of additional features from state-of-the-art researches

Feature Name	Accuracy
Structural	61.15%
Indicator	53.95%
Contextual	50.69%
Lexical	59.27%
Syntactic	50.72%
Prompt Similarity	54.79%
Word Embedding	49.46%
All Features	**71.08%**

5. CONCLUSIONS

After all the experiments, we have done to detect and classify the argument component, we found that 79.96% of accuracy was achieved by implementing all features set. We defined 68 sub-features which were summarized into 8 categories of features: they were structural, lexical, indicator, contextual, syntactic, word embed-ding, prompt similarity, and discourse features. We found that structural features were the best feature that had the most

significant impact to the system's accuracy, which obtained 77.83% accuracy. The other significant features are contextual and lexical, with the accuracy of 63.10% and 61.06%.

The most significant features combination was the combination of all features without discourse features. This combination obtained 79.98% accuracy, which was higher than the total accuracy of all features. The combination of all features without structural performed the lowest accuracy, so that we conclude that structural features was the most significant feature while discourse features was not. Besides, the combination of 3 (three) structural, contextual, and lexical features also performed a significant accuracy, which was 77.87%. Features proposed by Stab and Gurevych (2014b) performed the highest accuracy, which was 76.32%. Each of experiment in comparing features classification could obtain more than 67% of accuracy. It means that each of the experiment could identify argument components for more than 67%.

Since the experiments showed that the most significant features were structural, contextual, and lexical, we concern to develop these groups for our next experiment. We also find that the data training in bigger number with various topics and characteristics will probably increase the accuracy of system. Besides, we also must define the other features or the other method that can help in differentiate the premise and claim further.

6. ACKNOWLEDGMENTS

This research work was supported by Bina Nusantara University and partly supported by research grant from Directorate General of Research and Development Reinforcement, Ministry of Research, Technology and Higher Education of the Republic of Indonesia.

7. REFERENCES

[1] Khatib, A., Wachsmuth, H., Matthias, Hagen, M., Kohler, J., and Stein, B. 2016. Cross-domain mining of argumentative text through distant supervision. In 15th Conf. Of the North American Chapter of the Association for Computational Linguistics (NAACL'16) (to appear). Association for Computational Linguistics, San Diego, CA, USA, 2016.

[2] Feng, V.W. and Graeme H. 2011. Classifying arguments by scheme. Proceedings of 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, pp. 987-996, 2011

[3] Habernal, I. and Gurevych, I. 2015. Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015), pp. 2127-2137, Lisbon, Portugal, 2015.

[4] Habernal.I. and Gurevych, I. 2016. Argumentation mining in user-generated web discourse. Computational Linguistics, in press.

[5] Lawrence, J. and Reed, C. 2015. Combining argument mining techniques. Proceedings of the 2nd Workshop on Argumentation Mining, Denver, Colorado, pp. 127-136, 2015.

[6] Lippi, M. and Torroni, P. 2015. Context-independent claim detection for argumentation mining. Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015).

[7] Moens, M.F. 2014. Tutorial Argumentation Mining. Belgium

[8] Moens, M.F., Boiy, E., Palau, R.M. and Reed, C. 2007. Automatic detection of arguments in legal texts. In Proceedings of the 11th International Conference on Artificial Intelligence and Law, ICAIL '07, pages 225-230, Stanford, CA, USA.

[9] Palau. 2008. Automatic argumentation detection. Project ACILA - Automatic Detection and Classification of Arguments in a Legal Case, Leuven, Belgium.

[10] Palau, R.M. and Moens, M.F. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In Proceedings of the 12th International Conference on Artificial Intelligence and Law, ICAIL'09, pp. 98-107, Barcelona, Spain, 2009.

[11] Peldszus, A. 2014. Towards segment-based recognition of argumentation structure in short texts. Proceedings of the First Workshop on Argumentation Mining, pages 88-97,Baltimore, Maryland USA, June 26, 2014.

[12] Peldszus, A. and Stede, M. 2013. Form argument diagrams to argumentation mining in texts: a survey. International Journal of Cognitive Informatics and Natural Intelligence Volume 7 Issue 1, January 2013 Pages 1-31.

[13] Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A., Robaldo, L., and Webber, B.L. 2007. The Penn Discourse Treebank 2.0 annotation manual. Technical report, Institute for Research in Cognitive Science, University of Pennsylvania.

[14] Stab, C. and Gurevych, I. 2014a. Annotating argument components and relations in persuasive essays. In: Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014), pp. 1501-1510, Dublin, Ireland, 2014.

[15] Stab, C. and Gurevych, I. 2014b. Identifying argumentative discourse structures in persuasive essays. In: Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), pp. 46-56, Doha, Qatar, 2014.

[16] Stab, C. and Gurevych, I. 2016. Parsing argumentation structures in persuasive essays. In: arXiv preprint, under review, April 2016. Germany: Technische Universität Darmstadt.

[17] Stab, C. and Gurevych, I. 2016. Recognizing the absence of opposing arguments in persuasive essays. In: Proceedings of the 3rd Workshop on Argument Mining held in conjunction with the 2016 Annual Meeting of the Association for Computational Linguistics (ACL 2016), p. 113-118, August 2016

[18] Stab, C. and Habernal, I. 2015. Detecting argument components and structures. In: Report of Dagstuhl Seminar on Debating Technologies (15512), Vol. 5, p. 32-32, 2016.

[19] Toulmin, S. E. 1958. The Uses of Argument. Cambridge University Press.

[20] Wachsmuth, H., Khalid A. and Stein, B. 2016. Using Argument Mining to Assess the Argumentation Quality of Essays. Germany: Bauhaus-Universität Weimar.

[21] Walton, D. 2009. Argumentation Theory: A Very Short Introduction. In book: Argumentation in Artificial Intelligence, pp.1-24.