Argumentation mining: classifying argumentation components with Partial Tree Kernel and Support Vector Machine for constituent trees on imbalanced persuasive essay

H A Kusmantini\textsuperscript{1}, I Asror\textsuperscript{2} and M A Bijaksana\textsuperscript{3}
Telkom University, Jl. Telekomunikasi 1 Bandung 40257, Indonesia

E-mail:\textsuperscript{1}haniepong@student.telkomuniversity.ac.id, \textsuperscript{2}iasror@telkomuniversity.ac.id, \textsuperscript{3}arifbijaksana@telkomuniversity.ac.id

Abstract. Argumentation Mining is a method that automatically identified an argument structure in a text document. The structure of this argument consists of several components that become very important to evaluate itself. This study builds a classification system of 3 classes of argumentation components on persuasive essays so that the data is multi-class using Tree Kernel which is part of the pre-processing and Support Vector Machine as a tool for grouping this text. This research is an adaptation of research conducted by Lippi and Torroni where they used 2 classes to get 74.6% precision, recall 68.4% and F1-score 71.4%, while this study used 3 classes of argument class and managed to get a precision value of 79%, recall 78% and F1-score 74% by using the Sampling Method to overcome the problem of the amount of data imbalance.

1. Introduction
Argumentation is an important aspect of writing skills, not just a foundation to convince audiences of new ideas towards a particular perspective but also has a role in decision making and analyzing the attitudes of different authors [1]. Various forms of persuasive essay writing make it difficult for the reader to know the relevant arguments for the information. By using Argumentation Mining that automatically recognizes arguments in text documents, which can examine the text and the components of it to find out the quality of the argumentation. This assumption had been supported by the latest findings in psychology, which confirms that even general tutorials effectively improve the quality of the written argument [2]. Therefore we need a system that can automatically identify the components arguments which readers can directly evaluate the arguments in a persuasive essay component.

Research conducted by Lippi and Torroni by using argumentation mining on persuasive essay data collected by Stab and Gurevych where almost all sentences are annotated as premise or claim, or both based on the argumentation component using Partial Tree Kernel as method that can capture the structure of the argumentation component in the parse tree representation into the rich set feature. By adopting research conducted by Stab and Gurevych which uses 10-fold validation in 90 essays and considers the Claim and Major Claim categories as class targets and the relations in the components of the argumentation. They obtained a number of precision 74.6% performance, 68.4% recall and F1-score 71.4%.

In this study, the classification of components arguments on essay persuasive was built by adapting the research conducted by Lippi and Torroni which consists of three classes, namely Major Claim, Claim, and Premise which is a form of sentence to support or attack claims so that this component becomes attractive to be examined so that the argumentation of the essay itself can be evaluated [2]. This study also uses the Tree Kernel to process the argumentation component into the parse tree representation in the feature input space that will be examined in the pre-processing process and then classified in the Support Vector Machine (SVM) classification model. This study also uses a sampling
method as a test to solve the problem of the number of unbalanced classes that will be used. The Support Vector Machine is chosen because it can classify a large amount of data with protection that does not depend on the number of features but has the potential to handle large feature spaces and also has high data generalization capabilities [4].

2. Methodology

Most of the methods used in [3] are methods that overcome problems that are based on classifiers using Machine Learning which usually depend on a high set of computed feature sets. This method is driven by the observation that argumentative sentences are often characterized by general rhetorical structures. Parse Tree is a tree that is usually built based on either a grammatical constituent relation (grammatical phrase) or a grammatical dependency relation. Therefore, this study used parse tree constituents. Nodes in parse tree constituents are labeled standard non-terminal for grammar in English. For example, S is for sentence, VP is a verb phrase, NP is a noun phrase, DT is a determinant and others [3]. For example parse tree in sentence form that is detected as claim:

![Figure 1. Parse Tree constituents containing claim phrases.](image)

The illustration in figure 1 is the structure of a sentence that can be very informative to detect arguments. The parse tree constituent is the ideal instrument for representing that information. Because the research builds a component argumentation detection system based on a tool for classification, the Support Vector Machine (SVM) aims to capture the argument components between parse trees through the Tree Kernel [3]. The Kernel method is a method commonly used in a variety of NLP, especially Tree Kernel which has successfully overcome various problems related to NLP.

Tree Kernel (TK) is designed in such a way as to measure the similarity between two trees, by evaluating the number of their common substructure (or fragments). Usually, a TK is to consider the subtree as the selected fragment, as well as specifying more complex fragment structures [3]. The Tree fragments may be associated with different features in the high-dimensional vector space where the j-th feature only counts the number of j-th tree fragments, this can be counted as dot products between two representations of different trees. The kernel engine is then defined by exploiting structured information in the parse tree encoded by the Tree Kernel $K(x;z)$ function:

$$K(T_x, T_z) = \sum_{n_x \in N_{T_x}, n_z \in N_{T_z}} \Delta(n_x, n_z)$$

Which $N_{T_x}$ and $N_{T_z}$ is a collection of nodes or vertices of two trees, and $\Delta(\ , \ )$ is the value of two nodes based on the definition of the fragment. In this study the Partial Tree Kernel is used because according to research done by Lippi and Torroni, the PTK which allows to collect the most common fragments of the tree of sentence, would be the subtree that may be present at the node to be
considered, and since the higher the number of fragments the higher the $\Delta$ is the value between the two nodes [5]. However, the proposed TK Framework allows for inclusion in feature vector representations that can enrich feature descriptions from the considered examples. In this case, the kernel used will be counted as a combination between kernels between feature vectors (linear, polynomial, gaussian or rbf) [3].

2.1. Sampling method

In this study the number or ratio of the argumentation component is unbalanced, where the number of premise and claim will allow more than the number of Major Claim that there will only be one major claim that is the sentence presenting the topic or attitude of the author on an essay so this method is used to overcome the problem. In this Sampling method, there are two types that can handle the problem of data imbalance, namely Under-sampling and Over-sampling.

In general, the Under-sampling method is a step to eliminate data from negative classes or major classes. Some methods that can be used for this under-sampling use, namely one that will be used in this study with a Random-Under Sampler is a method that reduces the major classes contained in a random dataset document. Then another method, namely NearMiss, is a method that selects data in major classes that are identical to the data in the minor class [6].

Then the Over-sampling method is a step that attempts to add some minor classes or positive classes. Some ways that can be used for over-sampling are one of them that will be used in this study with Random-Over Sampler which selects data randomly in minor classes that will be duplicated [6]. Then the method of SMOTE (Synthetic Minority Oversampling Technique) which makes synthetic data that is replication data from minor class data [7].

2.2. Support Vector Machine

Support Vector Machine (SVM) was first introduced by Vapnik in 1992 as a harmonious series of superior concepts in the field of pattern recognition. SVM is a learning machine method that works on the principle of Structural Risk Minimization (SRM) with the aim of finding the best hyperplane that separates two classes in space input. SVM tries to find hyperplane by maximizing high generalization for future data [8].

Classification problems can be translated by finding a line (hyperplane) that separates between the two groups. The hyperplane is the best separator between the two classes can be found by measuring the margin of the hyperplane and finding its maximum point. A margin is a distance between the hyperplane and the closest pattern of each class. The closest data point is called a support vector. The solid line in figure 2 shows the best hyperplane, which is located right in the middle of the second data point, while the red and yellow points in the black circle are support vector. The effort to find the location of this hyperplane is the core of the learning process in SVM.

In this research, an SVM classification engine that exploits the kernel in the parse tree constituency is built, then the computed kernel above the parse tree, combined with the
computed kernel above the feature vector that is polynomial and Gaussian kernel (RBF), is used to train SVM groups in a set of examples labeled. The learning process in SVM to find support vector and only depends on dot product from the data in the feature space, but based on kernel parameters [8, 9].

2.3. Data
The data is a persuasive essay corpus on https://www.ukp.tu-darmstadt.de/. The Corpus contains an explanation of the components of the argument specifically, including Major Claim, Claim, and Premise with an argumentation relation of support and attack. A directed relation may apply on major claim with a claim, claim with a premise, or premise with another premise.

| Label     | Sentence                                      |
|-----------|-----------------------------------------------|
| Claim     | Competition can effectively promote the development of economy. |
| MajorClaim| We should attach more importance to cooperation |
| Premise   | The winner is the athlete but the success belongs to the whole team |

In total, the corpus contains 90 essays including 1,527 sentences consisting of argumentation components that are 90 of major claims, 423 of Claim, and 1,014 of Premise so that the data are imbalance for the classification system. In table 1 is the input form of the data used before being processed into the classification model, this research also does the work only on the classification of the text argument component without seeing the relation of each component of the argumentation.

3. Experiments
In the first part, the system will receive input from the argumentation data component. Then the data will go through the pre-processing stage where the argumentation component data will be converted into a parse tree representation form obtained using Stanford CoreNLP 3.5.04 software which provides complete syntactic analysis, including constituent representation and dependencies, based on probabilistic parsers [10].
In a study conducted by Moschitti using a value of 30 in the partial tree kernel feature space, whereas in this study using the input space in the Partial Tree Kernel feature space is 16, 32, and 64 to measure the optimal feature space for classification systems. Then the data is divided using the K-fold Validation method which is a technique of selecting a classification model where the data is broken down into k sections, then one (or more, for not a number of is taken as testing data, while the other is used as training data. Because the data used is multi-class so this study uses the One-against-all parameter to build a binary SVM model that compares one class to another [4]. In the Training process, the SVM model uses Polynomial and Gaussian (RBF) kernel parameters because this kernel model has a simple form and can be used in high-dimensional.

In this study, the value of Precision is needed which is an the accuracy of positive values that are well predicted compared to positive values (TP) plus the number of negative values predicted as positive values (FP), Precision = TP/TP+FP. Recall value is a calculation of the positive value predicted correctly (TP) compared to all positive values plus the number of positive values predicted as negative values (FN), Recall=TP/TP+FN. Then, the value of Accuracy to present the extent to which a classification system correctly compares all data accuracy = \( \frac{TP + TN}{P + N} \) This study also uses the F1-Score value as a harmonic average of precision and recall values where the two measurements are the gold standard of all performance tests on the classification model [8]. This research examines the classification system which is divided into 3 scenarios to solve data-balanced problems, namely:

(i) The first scheme of the SVM classification system uses original class data or imbalance class data which is the amount of argumentation component data in accordance with that in the corpus.
(ii) The next scheme of the classification system will use the Under-sampling method, namely Random-UnderSampler and NearMiss to reduce the number of major classes.
(iii) The final scheme, the classification system will use the Over-sampling method, namely the Random-OverSampler and SMOTE methods to add minor classes.

4. Results
Based on the evaluation method used, namely K-fold Validation, this study uses the value K=10 by adapting the classification engine built by Stab and Gurevych because the evaluation scores by accumulating the confusion matrices of each fold into one confusion matrix using kernel function parameters in the SVM classification system, and handling imbalance data with the sampling method. Optimal results of all tests performed with precision, recall, and F1-score values are displayed in Table 2.

| Scheme | Feature PTK | Sampling        | Kernel SVM | Accuracy | P   | R   | F1  |
|--------|-------------|-----------------|------------|----------|-----|-----|-----|
| 1      | 32          | -               | polynomial | 0.625    | 0.58| 0.62| 0.55|
| 2      | 64          | RandomUnderSampler | polynomial | 0.737704918 | 0.79| 0.74| 0.68|
| 3      | 64          | SMOTE           | polynomial | 0.778688525 | 0.79| 0.78| 0.74|

For the first scheme using original class data or imbalance class data can be seen from the accuracy results in figure 5, that the accuracy value is smaller than the results in figure 3 and 4 because the number of MajorClaim classes is too small compared to the number of other classes, besides that, the SVM kernel parameter, which is polynomial, can get the performance seen from the highest accuracy value of 62.5% and high F1-score compared to the RBF kernel.
In the third scheme in figure 4, the higher the parameter value of the PTK feature, the higher the accuracy value obtained. Besides that, in table 2 number 3 is the performance results obtained from this test are the accuracy of 77.86%, precision 79%, recall 78%, and F1-score 74% in polynomial kernels that get better performance compared to the RBF kernel. This is because the RBF kernel is stationary while the polynomial kernel is homogeneous and in this data homogeneous kernels can adjust better than stationary. Judging from the comparison between the two oversampling methods used that. Random Over-Sampling has a comparison that is not so far from the results obtained by using SMOTE whose accuracy value is better this is because SMOTE selects a new class that is randomly selected in the minor class then selects the five closest neighbors of the minor class in the training process.

5. Conclusion
After evaluation and analysis, it can be concluded that in the feature set parameter 64 the SMOTE method is used to handle the imbalance class problem and with the highest performance where the SVM kernel is polynomial which has reached 77.86% accuracy, precision 79%, recall 78%, and F1-score 74% for the 3 classes of argumentation components without seeing the relation of each component with the regenerated feature set to be a benchmark on the dimensions applied to the classification model. It is considered capable of adapting the research conducted by Lippi and Torroni where they obtained 74.6% precision, 68.4% recall and produced a score of 71.4% with 2 MajorClaim and Claim classes.
In this study also shows that the rich or high feature set, 32 and 64, has an accuracy value which is better than the low feature set but can still be used for both, namely the low feature set can use Over-sampling remembering the method this adds a minor class to the data, so it makes this method better than under-sampling. But the possibility of overfitted on the model that has been formed is relatively high considering that feature sets that can be generated can be even higher, so it is recommended to use optimization to measure the best feature set in this classification. Then for further research, it is recommended to apply a method that handles the problem of the amount of other imbalance data, to increase the possibility of this test getting better performance.

References
[1] Christian S and Iryna G 2014 Identifying argumentative discourse structures in persuasive essays EMNLP 2014 pp 46–56
[2] Christian S and Iryna G Annotating Argument Components and Relations in Persuasive Essays COLING 2014 pp 1501–1510
[3] Marco L and Paolo T 2015 Context-Independent Claim Detection for Argument Mining 2015 15 pp 185-191
[4] Suyanto 2017 Data Mining untuk Klasifikasi dan Klasterisasi Data (Bandung: INFORMATIKA) pp 115–246
[5] Alessandro M 2006 Efficient convolution kernels for dependency and constituent syntactic trees 2006 4212 pp 318–329
[6] Chung L A Y 2004 The effect of oversampling and undersampling on classifying imbalanced text datasets 2004 The University of Texas at Austin
[7] Nitesh V C et al 2002 SMOTE: synthetic minority over-sampling technique 2002 J. of artificial intelligence research 16 pp 321–357
[8] Anto S N, Arif B W and Dwi H 2003 Support vector machine teori dan aplikasinya dalam bioinformatika 2003 Kuliah Umum Ilmu Komputer.com
[9] Alexandre K 2014 SVM Tutorial url: https://www.svm-tutorial.com/2014/11/svm-understanding-math-part-1/.
[10] Christopher M et al 2014 The stanford CoreNLP natural language processing toolkit 2014 Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations pp 55–60