Solving the Imbalanced and Limited Data Labeled for Automated Essay Scoring using Cost Sensitive XGBoost and Pseudo-Labeling

Marvina Pramularsih[1], Mardhani Riasetiawan[2]*
Department of Computer Science and Electronics
Faculty of Mathematics and Natural Sciences
Universitas Gadjah Mada, Yogyakarta, Indonesia

***Abstract*—There are two main problems on forming the Automatic Essay Scoring Model. They are the datasets having imbalanced amount of the right and wrong answers and the minimal use of labeled data in the model training. The model forming based on these problems is divided into three main points, namely word representation, Cost-Sensitive XGBoost Classification, and adding unlabeled data with the Pseudo-Labeling Technique. The essay answer data is converted into a vector using the trained word vector fastText. Furthermore, the classification of unlabeled data was carried out using the Cost-Sensitive XGBoost Method. The data labeled by the classification model is added as training data for the new classification model form. The process is carried out iteratively. This research is about using the combination of Cost-Sensitive XGBoost Classification and Pseudo-Labeling which is expected to solve the problems. For the 0th iteration, the dataset having a ratio of the amount of "right" labeled data with the amount of "right" labeled data is close to 1, in other words a balanced dataset or a ratio that is more than 1 produces a model with better performance. Thus, the selection of training data at an early stage must pay attention to this ratio. In addition, the use of the Hybrid Method on these datasets can save labeled data 56 times compared to the AdaBoost Method. Hybrid model is able to produce F1-Measure more than 95.6%, so it can be concluded that the Hybrid Method, which combines the XGBoost and Pseudo-Labeling Cost-Sensitive Classification with Self Training, is able to overcome the problem of unbalanced datasets and data limited label.**

Keywords*—Imbalanced data; limited labeled data; automated essay scoring; cost sensitive XGBoost; pseudo-labeling

I. Introduction

Pusat Asesmen dan Pembelajaran (PUSMENJAR), Ministry of Cultural, Education, and Research Technology, Republic of Indonesia conducts a mapping program of educational attainment to monitor the quality of education nationally or locally, called AKSI (Asesmen Kompetensi Siswa Indonesia). Learning evaluation is carried out based on the results of the questions tested on students. PUSMENJAR distinguishes question packages into two, those are literacy and numeracy. There are eight packages of literacy questions and eight packages of numeracy questions that will be used to simulate the exam. In the sixteen question packages, there are six types of questions, namely true or false, check box, matching, sorting, multiple choice, and essay. For the process of correcting true or false, check boxes, matching, sorting, and multiple choice can be done by matching the answer keys. However, the process of correcting essay answers cannot always be done by matching an answer with the answer key. In addition, essay answers scoring manually takes longer than answers for multiple choice questions and short answers [1]. Therefore, we need models for scoring essay answers automatically.

Herwanto et al. [2] sees correction of the essay answer as a classification of a true or false answer using AdaBoost Classification Method. Another Automated Essay Scoring (AES) model has begun to be researched to develop a model for automatically correcting essay answers in Indonesian [3]. However, the models that have been developed have not paid attention to the effect of the large number of manually labeled data on model performance and have not paid attention to whether the dataset is a dataset that has amount of the right and wrong labeled data and is balanced or not.

The exam simulation of high school level questions obtained ten datasets of essay answers. After manual labeling by experts, it is known that the dataset is an imbalanced dataset between correct and false answers. Therefore, a classification algorithm which is capable of handling imbalanced data characteristics is needed. Fernandez et al. [4] and He and Ma [5] state that the approach taken to solve imbalanced data problems is divided into three points, namely methods at the data level, methods at the algorithm level, and methods at the hybrid level. The method at the algorithm level is the easiest method to apply [6]. Wang et al. [7] and Xia et al. [6] classify imbalanced data using the Cost-Sensitive XGBoost method. The use of the XGBoost algorithm is due to the fact there are many teams wining the competition using this algorithm [8]. In addition to the imbalance problem in the dataset, there is another problem that is forming a model with good performance with less training data than the current number of labeled data. The number of labeled data provided is currently around 6,000 data and will be used to correct approximately 330,000 uncorrected answer data. This problem can be overcome, one of which is by implementing Pseudo-Labeling with Self Training as was done by Babakhin et al. [9].

*Corresponding Author.

II. Automated Essay Scoring

Research on Automated Essay Scoring (AES) began since Page [10] conducted research on essay assessment using computers. AES research on answers to essays in Indonesia has begun to be developed from various points of view. One of them views the problem of essay assessment as a problem of classifying right or wrong answers [2], [3].

Herwanto et al. [2] conducted research on the AES Model for answers to essays in Indonesia. The dataset used is three datasets of student answers from the Program for International Student Assessments (PISA). The word representation used is Bag-of-Words (BoW) and character ngrams. The classification algorithm used is Adaptive Boosting (AdaBoost). The AES model formed has a F1-Score of 97.69% for the Machu Picchu dataset, 67.2% for the jacket dataset, and 71.74% for the bicycle dataset. Riasetiawan et al. [3] conducted research for essay answers on clustering and classification. The dataset used is a dataset of essay answers from the Indonesian Ministry of Education and Culture (Kemendikbud). The clustering algorithm used is K-Means, while the classification algorithm used is Convolutional Neural Network (CNN). Prior to clustering and classification, the word representation stage was carried out using GloVe. The answer classification model yields an accuracy above 85%.

In supervised learning, not all labeled datasets are balanced datasets. To solve the problem of unbalanced datasets, it can be done using data sampling methods [11], cost-sensitive algorithms [6], or a combination of data sampling methods and cost-sensitive algorithms [12]. The use of the data sampling method has weaknesses, namely, in addition to choosing a suitable classification method for the dataset, there is a need for further analysis of what data sampling method is more suitable for the dataset before entering the classification model training stage. In a cost-sensitive algorithm, there is no need to add these steps, so this method is the easiest of the other two methods to be applied to an unbalanced dataset.

Xu et al. [11] conducted research on the classification of sentiments and emotions on an unbalanced dataset. By using the Support Vector Machine (SVM) and Word Embedding Compositionality with Minority Oversampling Technique (WEC-MOTE), it can increase the precision by 29.3%. Xia et al. [6] conducted research on the classification of borrowers in peer-to-peer lending. By using the Cost-Sensitive XGBoost (CSXGBoost), the highest Area under the ROC Curve (AUC) value was obtained when compared to all trials compared by Xia et al. [6], which is 74.85%. The use of cost-sensitive is the easiest way to deal with unbalanced datasets [6]. Le et al. [12] conducted research on the classification of bankruptcy companies. Using CBoost and SMOTE-ENN, the highest AUC values were obtained from all the trials compared by Le et al. [12], which is 87.1%. Pseudo-Labeling is a technique used to increase the amount of labeled data by utilizing the unlabeled data that is owned. There are several learning algorithms used in Pseudo-Labeling, namely Self-Training [9], Co-Training [13], and Cluster-then-label [14]. In the Co-Training algorithm, the features used must be divided into two. Meanwhile, in cluster-labeling, there is a clustering stage before labeling. The clustering stage in the Cluster-then-label algorithm has its own challenges, namely making a good cluster. Based on the three algorithms, namely Self-Training [9], Co-Training [13], and Cluster-then-label [14], the Self-Training algorithm is the easiest algorithm to be implemented.

III. Hybrid Method

This research uses eight pairs of datasets, namely eight labeled datasets and eight unlabeled datasets [15], shown in Table I. Each dataset uses the .xlsx format. The datasets are from the PUSMENJAR. The datasets are the answers to the simulation of AKSI questions. For each pair the dataset will be analyzed and the model with the best performance is sought based on the experiments carried out.

Data pre-processing is necessary being used in the next process. The pre-processing consists of Lower-Case Folding, Filtering, and Tokenization, fastText pretrained word embedding for Indonesia. Pre-trained word vectors are trained on Wikipedia using fastText. The model is trained using skip-gram, dimension 100, subwords size 3-6 characters, epoch 5, and learning rate 0.05. To increase the chances that the machine can produce the best labels, the best-performing classification model is needed before it is used to predict answers on the unlabeled dataset. This is done by training the model with several combinations of different parameter values and selecting the model with the highest F1-Measure. Fig. 1 shows the Cost-Sensitive XGBoost modeling development for the research.

TABLE I. Dataset Description

Dataset		**Description**		
	Correct Label	***Wrong Label***	***Unlabeled***	***Total***
A	3.371	2.065	334.301	339.737
B	1.972	*3.824*	237.208	243.004
C	2.298	*3.525*	336.266	342.089
D	1.072	*4.716*	334.082	339.870
E	757	*5.046*	333.094	338.897
F	1.235	*4.605*	333.741	339.581
G	140	*5.566*	333.574	339.280
H	393	*5.288*	332.407	338.088

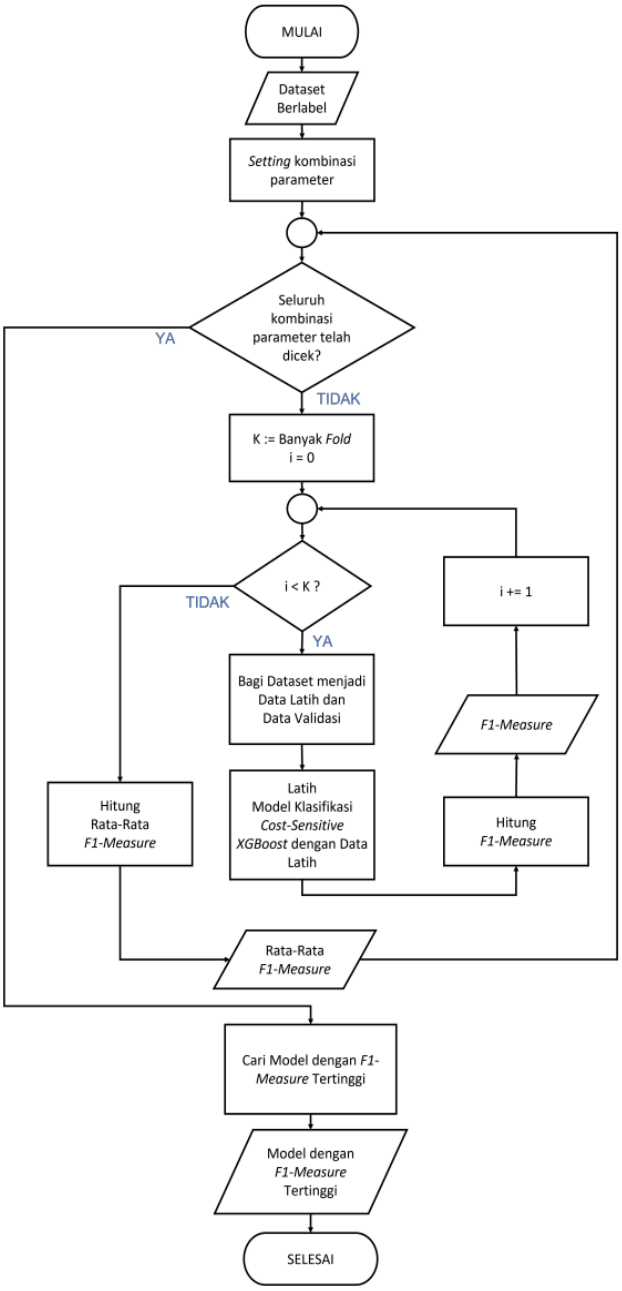

Fig. 1. Cost-Sensitive XGBoost.

IV. Result and Discussion

Pseudo-Labeling (Fig. 2) testing is seen based on two things, those are the amount of initial data and the amount of data added per iteration. This experiment is repeated until the third iteration. Fig. 3 (a)-(h) are the results of testing the initial data for Pseudo-Labeling. Based on Fig. 3, it is found that in the 0th iteration, the F1-Measure model with N0 = 100 is always lower than the model with more initial data. However, after iterating until the third iteration the F1-Measure of the model increases. On the Dataset 2, 4, 5, and 6, the F1-Measures of the models are able to exceed the models using a bigger amount of manual labeled data. In the Dataset 7, F1-Measure of the model with N0 = 100 is 0%. Although, it has been done for some iterations, but cannot improve the F1-Measure of the model. In this case, the number of "true" labeled data is 1, while the "false" labeled is 99. For the extreme case of the Dataset 7, the Pseudo-Labeling technique has not been able to improve the F1-Measure of the model. Furthermore, the amount of additional data tested are 50, 100, 150, and 200. Fig. 4 (a)-(h) are the results of testing additional data for Pseudo-Labeling using 100 labeled data as initial data. Based on Fig. 4, the selections of the amount of additional data per iteration that have been tried always increase F1-Measure of the models when compared to F1-Measure of the models in the 0th iteration, except for the Dataset 7 which only has 1 "true" labeled data and 99 "false" labeled data. Furthermore, the selection of additional data which is 50 (half of the number of initial data) is always lower than the others (equal to or higher than the initial data).

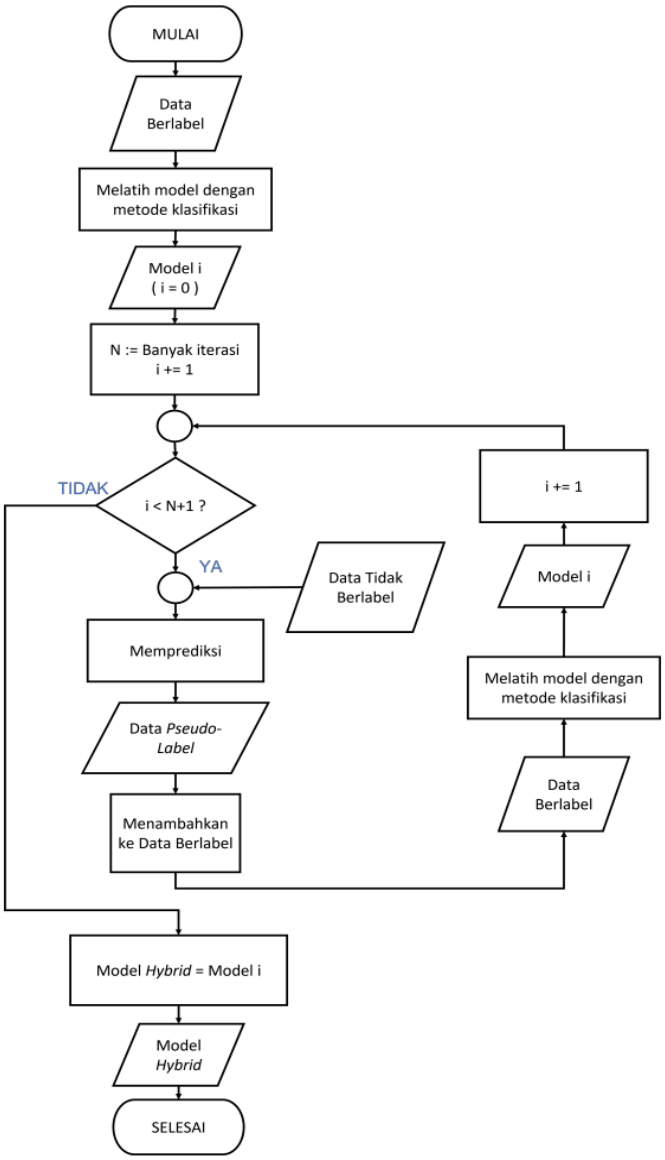

Fig. 2. Pseudo-Labeling with Self-Training.

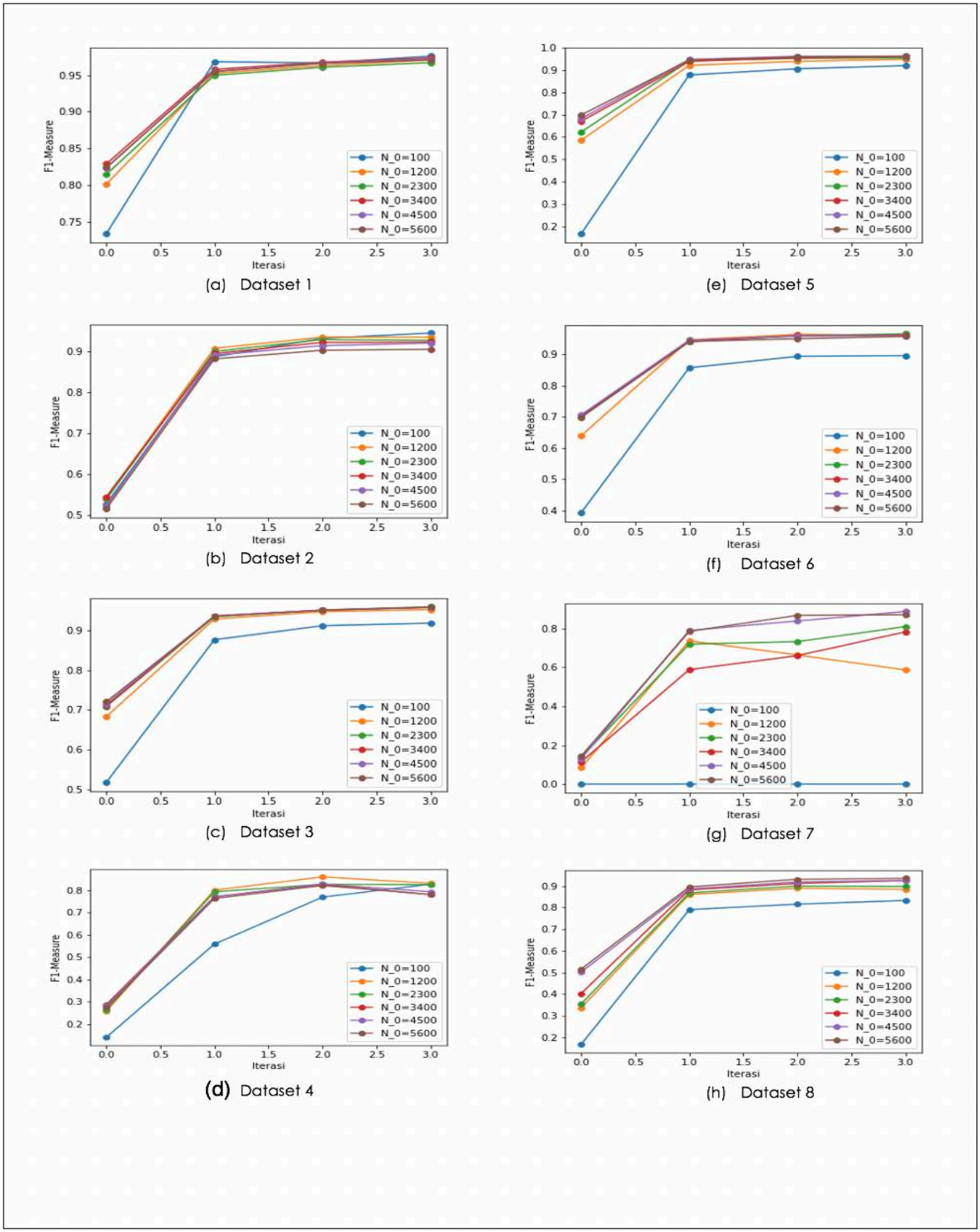

Fig. 3. Pseudo-Labeling F1-Measure Result..

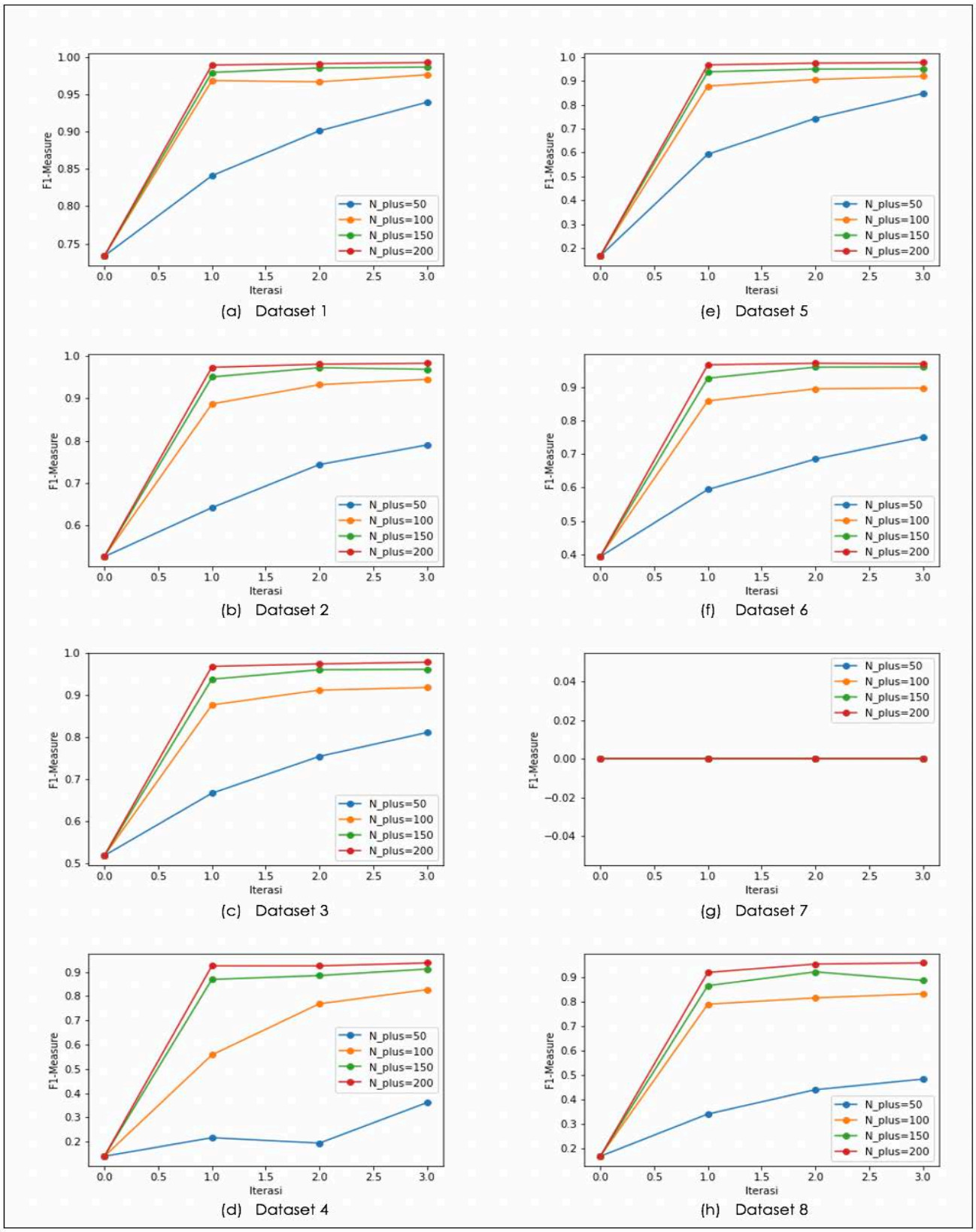

Fig. 4. Performance Comparisons for Cost-sensitive XGBoost.

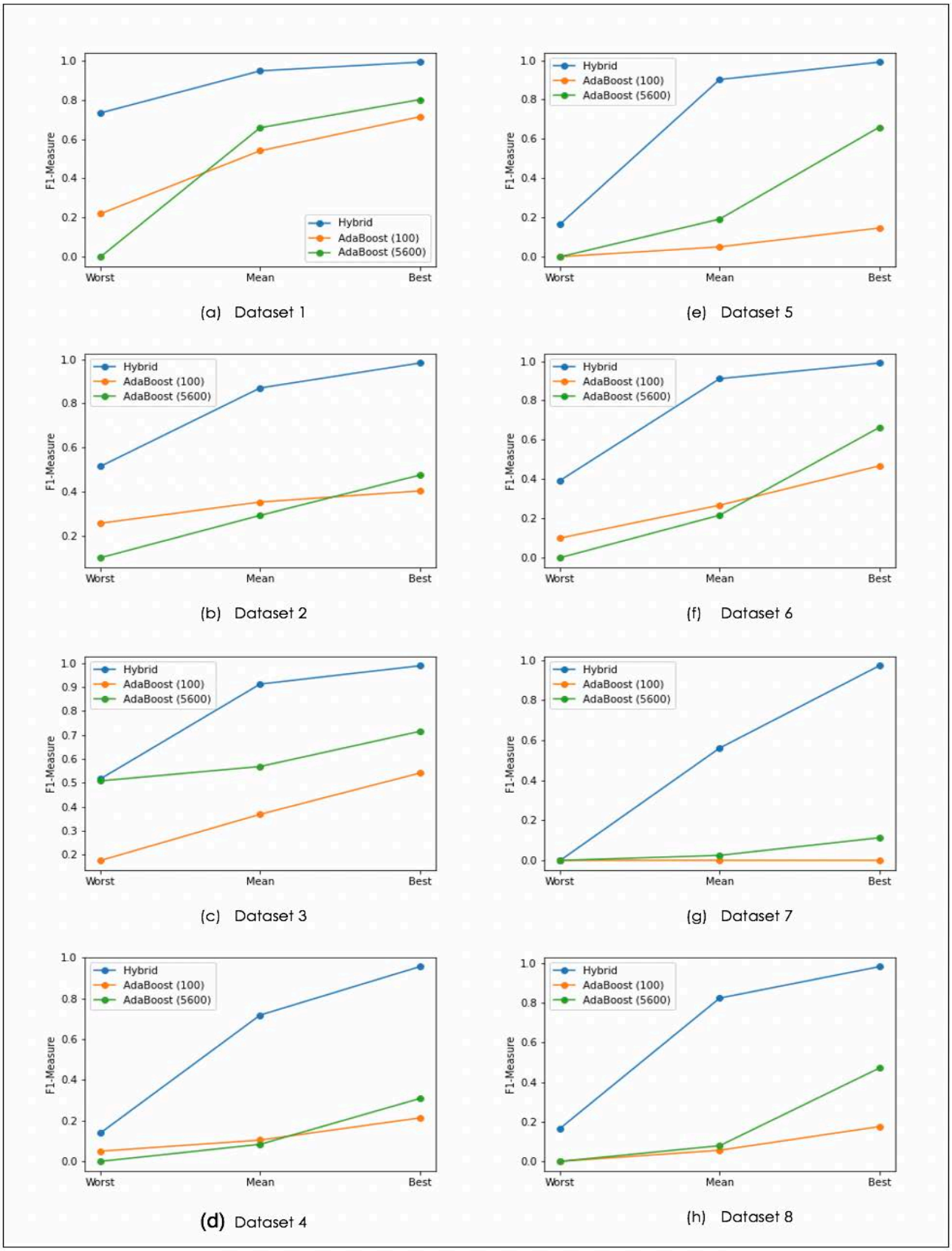

Fig. 5. Performance Comparisons for Hybrid and AdaBoost.

Fig. 5 (a)-(h) are graphs comparing three models, namely the Hybrid Model using 100 labeled data, the AdaBoost Model using 100 labeled data, and the AdaBoost Model using 5600 labeled data. In this comparison, three F1-Measure values are compared for each case, namely the best, mean, and worst F1-Measures that can be obtained from the use of these algorithms.

Based on Fig. 5 (a)-(h), it is found that the best and mean F1-Measure of the Hybrid Model is always higher than the two AdaBoost Models. Overall, the worst F1-Measure of the Hybrid Model is also higher than the two AdaBoost Models. There is only one worst F1-Measure value of the Hybrid Model which is lower than the AdaBoost Model using 5600 labeled data, namely the Dataset 3, but the difference in F1-Measure values is less than 6%. By saving 5500 data labeling, the difference between the F1-Measure values can be neglected. Overall, even though one of the AdaBoost Models already uses 56 times more data, it still cannot compete the F1-Measure of the Hybrid Model.

V. Conclusion

For the 0th iteration, the dataset having a ratio of the amount of "right" labeled data with the amount of "right" labeled data is close to 1. In other words, a balanced dataset or a ratio that is more than 1 produces a model with better performance. Thus, the selection of training data at an early stage must pay attention to this ratio. In addition, the use of the Hybrid Method on these datasets can save labeled data 56 times compared to the AdaBoost Method. The positive class weight parameter has no effect on the performance of the resulting model. The Pseudo-Labeling process with Self Training is able to handle the problem of limited training data, except for the Income Dataset Residents who have F1-Measure with a value of 0% both before and after the Pseudo-Labeling process with Self Training. Hybrid model which is able to produce F1-Measure more than 95.6%, so it can be concluded that the Hybrid Method combines the XGBoost and Pseudo-Labeling Cost-Sensitive Classification with Self Training is able to overcome the problem of unbalanced datasets and data limited label.

Acknowledgment

This work is supported by Pusat Asesmen dan Pembelajaran (PUSMENJAR), Ministry of Cultural, Education, and Research Technology, Republic of Indonesia for providing datasets. The paper is part of the Master Thesis by the 1st author in Magister of Computer Science, Faculty of Mathematic and Natural Sciences, Universitas Gadjah Mada.

References

[1] Valenti, S., F. Neri, and A. Cucchiarelli, An Overview of Current Research on Automated Essay Grading. *Journal of Information Technology Education: Research.* 2: 319-330, 2017.

[2] Herwanto, G. B., Y. Sari, B.N. Prastowo, M. Riasetiawan, I.A. Bustoni, and I. Hidayatulloh, UKARA: A Fast and Simple Automatic Short Answer Scoring System for Bahasa Indonesia. *Proceeding Book of 1st International Conference on Educational Assessment and Policy*. 2: 1-8, 2018.

[3] Riasetiawan, M., B.N. Prastowo, I. Novindasari, and N.J. Aisyiah, Automatic Scoring System for Essay Answer Data using Computational Approach: Clustering and Convolutional Neural. *Prosiding 1st National Conference on Educational Assessment and Policy (NCEAP 2018).* 1: 89-96, 2018.

[4] Fernández, A., S. García, M. Galar, R.C. Prati, B. Krawezyk, and F. Herrera, *Learning from Imbalanced Data Sets.* Switzerland: Springer, 2018.

[5] He, H. and Y. Ma., Imbalanced Learning: Foundations, Algorithms, and Applications. New Jersey: John Wiley & Sons, 2013.

[6] Xia, Y., C. Liu, and N. Liu, Cost-Sensitive Boosted Tree for Lean Evaluation in Peer-to-Peer Lending. *Electronic Commerse Research and Applications*, 2017.

[7] Wang, C., C. Deng, and S. Wang, Imbalance-XGBoost: Leveraging Weighted and Focal Losses for Binary Label-Imbalanced Classification with XGBoost. *Pattern Recognition Letters.* 136: 190-197.

[8] Chen, T., and C. Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Agustus 2016. 785-794, 2020.

[9] Babakhin, Y., A. Sanakoyeu, and H. Kitamura, Semi-Supervised Segmentation of Salt Bodies in Seismic Images using an Ensembles of Convolutional Neural Networks. *Pattern Recognition 41st DAGM German Conference*. Dortmund, Germany. 10-13 September 2019. 218-231, 2019.

[10] Page, E. B, The Imminence of Grading Essays by Computer-25 Years Later. *Computers and Composition.* 10(2): 45-58, 1993.

[11] Xu, R., T. Chen, Y. Xia, Q. Lu, B. Liu, and X. Wang, Word Embedding Composition for Data Imbalances in Sentiment and Emotion Classification. *Cognitive Computation.* 7: 226-240, 2015.

[12] Le, T., M.T. Vo, B. Vo, M.Y. Lee, and S.W. Baik, A Hybrid Approach using Oversampling Technique and Cost-Sensitive Learning for Bankruptcy Prediction. *Complexity.* 2: 1-12, 2019.

[13] Didaci, L., G. Fumera, and F. Roli, Analysis of Co-Training Algorithm with Very Small Training Sets. *Structural, Syntactic, and Statistical Pattern Recognition.* Hiroshima, Japan. 7-9 November 2012. 719-726, 2012.

[14] Peikari, M., S. Salama, S.N. Mozes, and A.L. Martel, A Cluster-then-label Semisupervised Learning Approach for Pathology Image Classification. *Scientific Reports.* 8(1): 7193, 2018.

[15] Pramularsih, M. Cost-Sensitive XGBoost and Psesudo-Labeling with Slef Training for Imbalanced Data and Few Labeled Data in Automated Essay Scoring, Master Thesis in Magister Program of Computer Science, Faculty of Mathematic and Natural Sciences, Universitas Gadjah Mada, Indonesia, 2021.