# INTISARI

## *COST-SENSITIVE XGBOOST* DAN *PSEUDO-LABELING* DENGAN *SELF TRAINING* UNTUK DATASET YANG TIDAK SEIMBANG DAN DATA BERLABEL YANG TERBATAS PADA *AUTOMATED ESSAY SCORING*

Oleh

MARVINA PAMULARSIH

19/448714/PPA/05797

Pusat Asesmen dan Pembelajaran (Pusmenjar) melalui Asesmen Kompetensi Siswa Indonesia (AKSI) mulai mengembangkan soal dengan tipe jawaban esai untuk melihat kompetensi siswa. Proses pengoreksian jawaban esai secara manual memerlukan waktu yang lama dan memakan banyak biaya. Oleh karena itu, perlu adanya *Automated Essay Scoring* (AES), sehingga mengurangi beban kerja pengoreksian secara manual.

Terdapat dua permasalahan utama pada Model AES yang akan dibentuk, yaitu dataset-dataset yang diperoleh mempunyai banyak jawaban benar dan salah yang tidak seimbang dan penggunaan data berlabel yang seminimal mungkin dalam proses pelatihan model. Pembentukan model berdasarkan permasalahan tersebut dibagi menjadi tiga poin utama, yaitu representasi kata, pembentukan model klasifikasi dengan *Cost-Sensitive XGBoost*, dan penambahan data tidak berlabel dengan Teknik *Pseudo-Labeling*. Data jawaban esai diubah menjadi vektor menggunakan *fastText pre-trained word vectors*. Selanjutnya, dilakukan klasifikasi untuk data tidak berlabel dengan menggunakan Metode *Cost-Sensitive XGBoost*. Data yang telah diberi label oleh model klasifikasi tersebut ditambahkan sebagai data training untuk membentuk model klasifikasi yang baru. Proses tersebut dilakukan secara iteratif. Penelitian ini mengusulkan penggunaan Metode *Hybrid* yang mengombinasikan Metode Klasifikasi *Cost-Sensitive XGBoost* dan *Pseudo-Labeling* dengan *Self Training* yang diharapkan dapat mengatasi kedua permasalahan tersebut.

Dengan Metode *Hybrid* yang mampu dihasilkan *F1-Measure* lebih dari $95.6\%$. Dengan kata lain, Metode *Hybrid* mampu mengatasi permasalahan dataset yang tidak seimbang dan data berlabel yang terbatas pada AES. Serta, penggunaan Metode *Hybrid* yang memperhatikan kedua permasalahan tersebut pada kedelapan dataset lebih baik dari Metode AdaBoost yang tidak memperhatikan kedua hal tersebut.

**Kata Kunci:** ***Automated Essay Scoring*****, FastText, Klasifikasi** ***Cost-Sensitive XGBoost*****,** ***Pseudo-Labeling***

# ABSTRACT

## COST-SENSITIVE XGBOOST AND PSEUDO-LABELING WITH SELF TRAINING FOR IMBALANCED DATA AND FEW LABELED DATA IN *AUTOMATED ESSAY SCORING*

By

MARVINA PAMULARSIH

19/448714/PPA/05797

Pusat Asesmen dan Pembelajaran (Pusmenjar) through Asesmen Kompetensi Siswa Indonesia (AKSI) starts to develop questions with essay answers to observe students' competencies. The process of manually correcting essay answers is time consuming and costly. Therefore, it is necessary to have Automated Essay Scoring (AES), thereby it can reduce the workload for manual corrections.

There are two main problems on forming the AES Model. Those are the datasets having unbalanced amount of the right and wrong answers and the minimal use of labeled data in the model training. The model forming based on those 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 called Hybrid Method. The method is expected to solve the problems.

By applying Hybrid Method, we can achieve F1-Measure more than $95.6\%$, it means Hybrid Method has capability to solve the problems. Those are imbalanced data and few labeled data in AES. It is also known that Hybrid method paying attention to the problems performs better than Metode AdaBoost which not pays attention to the problems.

**Keywords: Automated Essay Scoring, FastText, Cost-Sensitive XGBoost Classification, Pseudo-Labeling**