28th Meeting of the Wiesbaden Group on Business Registers
The Hague, The Netherlands, 2 - 6 October 2023

Annisarahmi Nur Aini Aldania, Sheerin Dahwan Aziz, Irma Damayanti, Jhonathan Putro Siahaan, Imam Satya Wedhatama
BPS - Statistics Indonesia

Session 4: Industrial Classification Systems: Treatments for the upcoming NACE or ISIC revision and other industry classification issuess

Artificial Intelligence for Predicting Indonesia Industrial Classification Code

Abstract

Keywords: Industrial Classification, Artificial Intelligence, Text Analytics, IndoBERT, Double Random Forest

Industry classification is a rule or principle for grouping a company based on its economic activities into specific classes. BPS – Statistics Indonesia published the Standard Classification of Indonesian Business Fields (KBLI) as a guideline for classifying companies. In the KBLI guidelines, each company can be classified into a five-digit number in the KBLI group according to details in the form of text data on the main activities and the main products produced. KBLI is arranged hierarchically, consisting of one letter digit for a category, two-digit numbers for main groups, three-digit numbers for groups, four-digit numbers for subgroups, and five-digit numbers for groups. From the point of view of machine learning or AI, KBLI can be seen as a multi-class classification problem so that a model can be formed that can be used to generate predictions for the five-digit KBLI group in a company based on the description of the main activities and main products produced by the company.

Multi-class classification tends to be more complex than two-class classification. One thing that affects the difficulty of prediction is the interaction pattern that becomes increasingly complicated between independent variables and response variables as the number of classes increases. The ensemble method is a method that can be used to solve multi-class classification problems. Compared to using a single tree, the ensemble method combines several single trees to make predictions, one of which is Double Random Forest (DRF). In addition, the development of Natural Language Processing (NLP) in Indonesian also presents new methods worthy of comparison, one of which is IndoBERT. Both models can predict the five-digit KBLI group of a business or company based on the main activities and products produced. Thus, reducing the manual matching process and can be used for quality assurance for existing business or company databases.