

## Using Big Data Tools and Machine Learning Techniques to Assign Classification of Individual Consumption by Purpose (COICOP) Categories

Organisation: Turkey Statistical Institute (TURKSTAT)  
Author(s): Abdulcebar On, Halil Zeybek, Ali Osman Bilgin,  
Kudret Elif Berkman, Mustafa Karamavus  
Date: 24.01.2022  
Version: 3.0

### **Background**

Consumer price index (CPI) is the change ratio in the price of basket of goods and services purchased by a group of consumers over time. CPI is measured by using product groups which are classified based on Classification of Individual Consumption by Purpose (COICOP) categories. In the process of producing CPI, each product which is collected by regional officer must be assigned to relevant COICOP category. This operation of labeling product data is conducted by applying a set of rules in a form of Python script or PL/SQL procedures. Limitation of this approach is that rules must be updated manually as new products are added to product dataset. Furthermore, scanner data from the private sector and web scraping data from web sites have been added as data sources for collecting product data. As a result of increase in number of collected products, updating rules manually become more problematic.

In order to overcome this limitation, a generic classifying system which is based on machine learning and deep learning algorithms should be developed. Benefit of using machine learning to classify related COICOP categories for each product is that after training the system with new products features including brand name and product name, it automatically learns how to classify new products without any direct modification on the system. By using big data technologies to store and process high volume of product data and machine learning algorithms to automatically classify products on COICOP categories, a more scalable and maintainable product classification system will be developed.

## Input Data

Input data set consists of the product data of companies in the food and non-alcoholic beverages and clothing and footwear sectors covered by the Consumer Price Index (CPI). The data set has been obtained from 3 different sources:

- Barcode data of companies in the relevant sectors (107 951 rows).
- Survey data collected by Regional Statistic Offices in Turkey (127 770 rows).
- By extracting data from websites with web scraping method (not used yet).

The data set contains 2 columns. The first one is “COICOP Classification Code (7 digits, 420 different classes)” which is a dependent variable for our models. The second one is “Text Data (product definitions written in Turkish)” which is an independent variable for our models.

## Data preparation

In this study, for all of the methods that were used, very similar data preprocessing steps were performed. On the other hand, no preprocessing was carried out in the Bidirectional Encoder Representations from Transformers (BERT) model. In data preparation step, punctuation marks within the text data were removed first. After that, all letters were converted to lowercase letters of the Turkish character set. This process is followed by the tokenization process. With this process each sentence is separated into the words that form them. Finally, stop words of the Turkish language in the NLTK library have been eliminated. During the preprocessing stage, detection and correction of spelling errors, stemming and n-gram operations were not performed yet.

## Training Data

After the data preparation stage, the training data to be inserted into the models (except BERT model) was prepared. The output data consists of a two-dimensional array. One line in the array corresponds to one product description. On the other hand, a cell in the array is a word in the product definitions called token. Since the word count in the product definitions are different from each other, the size of the inner arrays is different from each other.

## Machine Learning Solution

In this study we have tried three machine learning models so far: Logistic Regression, Support Vector Machine, and Naive Bayes. All this models were implemented with the popular Python machine learning library “scikit-learn”. The dataset is split into training and test sets (%85 training, %15 testing) by using 42 for random state. TFIDF, short for term frequency–inverse document frequency, was used in a pipeline for all models. Due the possibility of less frequent words slowing down the model formation process and causing incorrect learning within the model, maximum word count was set to 4000, in order to mitigate this possibility. Finally, accuracy, precision, recall, and f-score values are calculated

by “using precision recall f-score support” and “accuracy score” from “scikit-learn” library as following:

| Model                  | Accuracy | Precision | Recall | F-score |
|------------------------|----------|-----------|--------|---------|
| Logistic Regression    | 0.96     | 0.94      | 0.93   | 0.93    |
| Support Vector Machine | 0.94     | 0.90      | 0.90   | 0.90    |
| Naive Bayes            | 0.92     | 0.83      | 0.79   | 0.80    |

We are also planning to implement other machine learning models such as Decision Tree, Random Forest, KNN and Multilayer Perceptron to compare the success and duration of the models on COICOP data set.

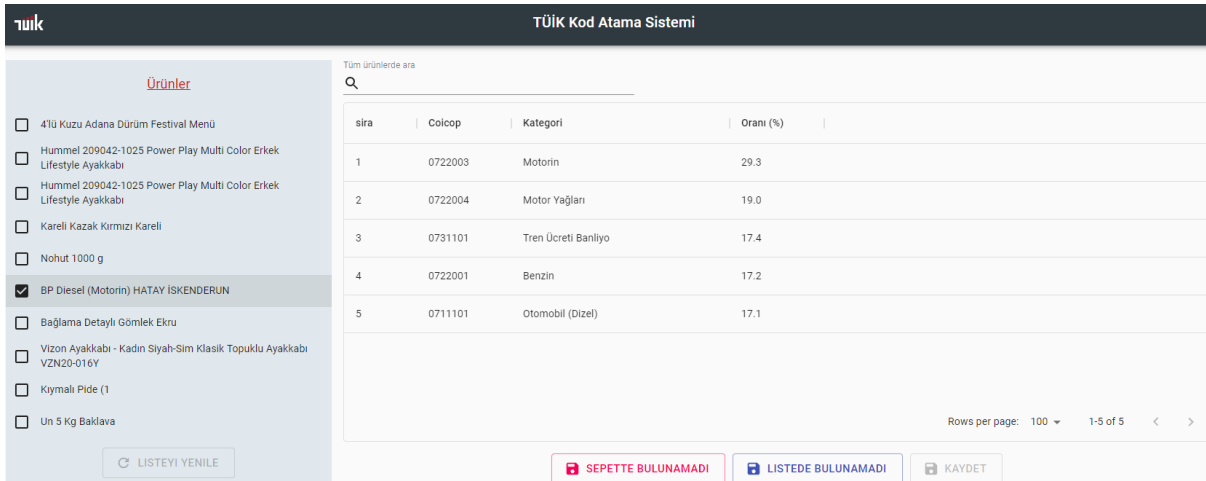
### Deep Learning Solution

In our deep learning approach, we have tried 3 different BERT (Bidirectional Encoder Representations from Transformers) based pre-trained models. To be able to try these models, we first chose 3 pre-trained models which are trained with Turkish corpus on HuggingFace website. We trained all these 3 models with our own labeled data on Google Colab environment and inspect their accuracy in terms of precision, recall and f1-score values. We have selected the best model in terms of these values and downloaded the best model to our local environment. By using Flask API with Python, we have served our model. Precision, recall and f1-score values of all these 3 models are shown below.

Table-1: Model Results

| MODEL                                | Precision | Recall | F1-Score |
|--------------------------------------|-----------|--------|----------|
| dbmdz/bert-base-turkish-cased        | 0,91      | 0,92   | 0,91     |
| bert-base-multilingual-cased         | 0,91      | 0,91   | 0,91     |
| dbmdz/bert-base-turkish-128k-uncased | 0,93      | 0,92   | 0,93     |

The average accuracy was used by combining the results of these 3 models. For this, it has been converted into a microservice using Python's Flask library. A code assignment screen that can use the service has been developed. On the code assignment screen, users can list the coicop estimates from the service by selecting the product names from web crawling.



| sıra | Coicop  | Kategori            | Oranı (%) |
|------|---------|---------------------|-----------|
| 1    | 0722003 | Motorin             | 29.3      |
| 2    | 0722004 | Motor Yağları       | 19.0      |
| 3    | 0731101 | Tren Ücreti Banlıyo | 17.4      |
| 4    | 0722001 | Benzin              | 17.2      |
| 5    | 0711101 | Otomobil (Dizel)    | 17.1      |

Figure-1: Product code assignment screen with average estimation of BERT models

### Labelling Products by Using CNN-based Pretrained Model (Resnet152)

Product images were obtained by using the Google search engine by web crawling method. Product images are sorted into food categories in coicop 7 digit. The separation process was carried out as follows. First, a Web crawling application was developed with the Java programming language. With this application, the food category defined in Coicop was queried via the Google image search engine and 300 of the listed images were saved in the relevant coicop category. In this way, 52130 pictures consisting of 134 categories were recorded in the system. The created data set is used for the Resnet152 model. After the model was trained, an f1 score of 89 percent was obtained.

This model was used for product images collected by statistical offices located in the regions. It was used to extract real pictures from the collected pictures. For example, the pictures in the rice category that could actually be rice were determined by the model.

### Next Steps

We developed a web scraping system that collects product data including price, product title, product description and unit of measurement from several different web pages. After improving the accuracy of the model iteratively by increasing the number of high quality labelled training data for each COICOP group, we will be able to daily assign COICOP to each product data by using the machine learning model. After increasing the model accuracy for COICOP 7 digits, we will develop a final model for COICOP 11 digits.

We are also planning to execute deep learning codes (Tensorflow/PyTorch) in a distributed environment to reduce the training time. In order to increase accuracy of the model, we are planning to combine our NLP-based deep learning solution and our Convolutional Neural Network (CNN) based (Resnet152) solution by using products images.

We are planning to extend this study to assign PRODCOM (Community Production) codes based on product definitions and evolve our system from a specific-purpose code assignment system to a more general-purpose one that can assign code for different domains.