TURKISH STATISTICAL INSTITUTE

14th OIC-StatCom Meeting

Modernization of National Statistical Offices through Digital Transformation

Ali Osman BİLGİN

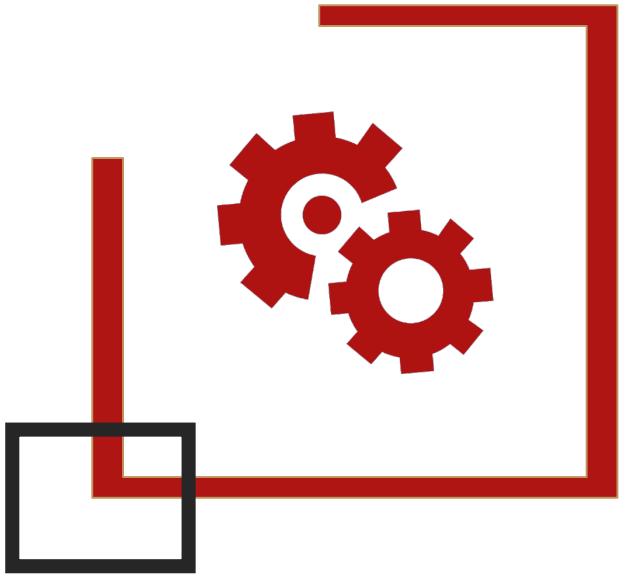
Head of the Artificial Intelligence and Data Analysis Group

The Artificial Intelligence and Data Analysis Group

The Artificial Intelligence and Data Analysis Group, which is part of the Data Governance Department, is doing a bunch of stuff to modernize national statistics offices through digital transformation.

With these efforts, processes that used to be done the old-fashioned way are now being done faster, more reliably, and more efficiently with:

- Artificial intelligence
- Big data technologies
- Automation systems



Career Orientation Survey: Maya Project

Career Orientation Survey: Maya Project Purpose

- The aim is to **systematically analyze** job advertisement texts on the Internet.
- The **occupational and skill information** extracted from job advertisements will be labeled according to international standards to make it **more reliable and comparable**.
- The main objective of the study is to **analyze current trends in the labor market and make predictions about future job trends.**
- The aim is to contribute **to strategic planning and policy development processes** for the public, academic, and private sectors.

Career Orientation Survey: Maya Project Goal

Automatic Classification of Occupations

- Labeling of occupations mentioned in job advertisements according to the ISCO-08 (International Standard Classification of Occupations 2008) system.

Skill Identification and Labeling

- Extraction and classification of skills and competencies in job postings according to the ESCO (European Skills, Competences, Qualifications and Occupations) system.

Guidance Reporting for Stakeholders

- Production of strategic reports for policymakers, universities, vocational training institutions, and employers.

Career Orientation Survey: Maya Project Outcome

- **Automation compared to manual classification:** Fully automated labeling will be provided without the need for experts to analyze job postings one by one.
- **Standardized data:** Data obtained through ISCO-08 and ESCO-compliant labeling will be comparable at the international level.
- **Contribution to labor force policies:** Educational institutions and decision-makers will be able to predict the skills that will be needed in the future.
- **Increased efficiency:** The system developed through academia-public sector collaboration will save time and human resources.
- **Knowledge accumulation:** The models and rules developed at the end of the project will be reusable in similar field studies.

Career Orientation Survey: Maya Project Method

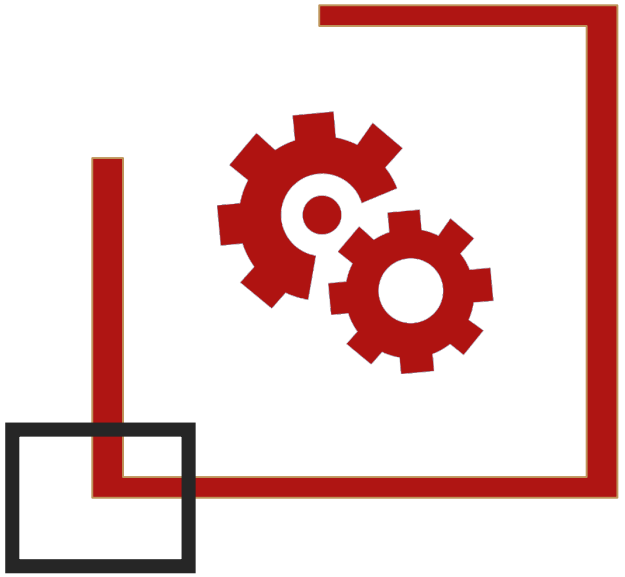
Text Processing:
Deep learning-based methods (e.g., **SBERT**) will be used to process job posting texts.
Raw texts will be divided into individual sentences using heuristic algorithms.

Occupation Labeling:
- Job postings will be classified using ISCO-08 codes.
- For this purpose, an embedding model based on all-MiniLM-L6-v2 was trained, and the **Multiple Negatives Ranking Loss function** was used.
- Similarities are calculated using a 124,000-line synthetic ISCO-08 dictionary, and the closest occupation is selected using the **MaxSim reordering method**.

Skill Labeling:
- The ESCO classification was used as a basis.
- **Process**:
  - Intuitive NLP techniques
  - Traditional classifiers **(Logistic Regression / XGBoost)**
  - Fine-tuned LLM-based automatic translation + semantic similarity calculation
- **Objective**: To extract and standardize the correct skills from dirty job posting data.

Career Orientation Survey: Maya Project Result

- This project will use a fully automated artificial intelligence system to analyze **occupations and skills based on job postings.**
- Thanks to **ISCO-08 and ESCO-compliant labeling**, the results will be comparable at the national and international levels.
- The outputs of the study:
  - **Forecast reports for policymakers**
  - **Skill needs analyses for educational institutions**
  - **Information on market trends for employers**
- The project, carried out with the support of **TÜBİTAK ARDEB 3005**, will be an important example of public-academic collaboration and will provide **strategic roadmaps** for the future of the labor market.



Artificial Intelligence-Supported Automation of Media Reports on TÜİK

Artificial Intelligence-Supported Automation of Media Reports on TÜİK Purpose

- Manually tracking and reporting news about the institution in the media is a very time-consuming process.

- With this project, the aim is to:
  - **Save manpower,**
  - **Standardize processes,**
  - Make analyses **fast, consistent, and repeatable.**

Artificial Intelligence-Supported Automation of Media Reports on TÜİK
Goal

Automatic News Topic Detection

- Automatic assignment of correct topic headings based on news content.

News Bulletin Creation

- Preparation of automatic news summaries and bulletins for daily/weekly reporting.

Author Recognition

- Automatic identification of the author or source of the news.

Positive/Negative Analysis of News

- Automatic classification of the sentiment orientation (positive, negative, neutral) of news.

Artificial Intelligence-Supported Automation of Media Reports on TÜİK Outcome

Reduced manual workload: Routine media scanning tasks performed by specialist staff will be automated.

Time and cost savings: More news can be processed in less time.

Standardized analysis: Consistent and systematic outputs will be obtained instead of subjective interpretations.

Focus on value-added work: Specialist staff will be able to focus on more strategic and creative tasks.

Artificial Intelligence-Supported Automation of Media Reports on TÜİK Method

Various deep learning and rule-based methods will be used.

Topic Assignment
- Machine translation models
- BART-based models *(e.g., facebook/bart-large-cnn)*
- Fine-tuned SBERT models *(e.g., all-MiniLM-L6-v2)*

Newsletter
- Rule-based system based on linguistic templates and rules

Author Recognition
Machine translation models
BERT-based models for NER *(e.g., dslim/bert-base-NER)*

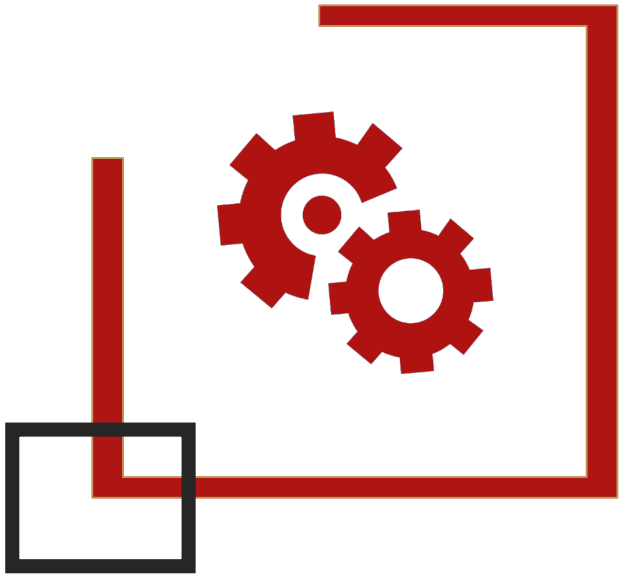
Positive/Negative Analysis
Machine translation + BART-based summarization
Transformer models for sentiment analysis *(e.g., distilbert-base-uncased-finetuned-sst-2-english)*

Artificial Intelligence-Supported Automation of Media Reports on TÜİK Result

The system to be developed within the scope of the project will offer a **multidimensional natural language processing infrastructure.**

Thanks to automatic news tracking and analysis:

- **Productivity will increase,**
- **Human errors will decrease,**
- **More strategic media management** will become possible.



Big Data Advanced Analytics Project – Web Scraping

Big Data Advanced Analytics Project – Web Scraping Purpose

- Price information related to workplaces in Turkey, job advertisements related to **the labor market, and price data for agricultural products and tractors** are systematically collected using web scraping techniques.

- The collected data is used to support **statistical data production processes and improve data quality within the framework of big data analytics**.

- The project automates **data processing and analysis processes to create a fast, reliable, and repeatable data infrastructure.**

Big Data Advanced Analytics Project – Web Scraping Goal

- **Job Posting Data Collection**
  - Job postings are automatically scraped from career websites in Turkey on a daily basis.
- **Price Data Collection**
  - Price data for agricultural products, tractors, and other workplaces is collected daily through web scraping.
- **Data Integration and Big Data Management**
  - The collected data is transferred to and managed in a big data environment using the Streamsets platform and Apache Kafka.
- **Analysis and Statistical Production**
  - The obtained data can be analyzed using Impala and contributes to statistical data production processes.

Big Data Advanced Analytics Project – Web Scraping Outcome

- **Automatic data collection:** 130,000 product and price details from 47 websites, 15,000 job listings from 14 career websites, and price data from 3 agriculture/tractor websites are collected automatically every day.
- **Optimization of data processing processes:** Data flow and integration are standardized using Apache Kafka, Streamsets, and HDFS.
- **Data quality and reliability:** Data quality is improved through regular and automatic control mechanisms.
- **Statistical decision support:** A reliable data infrastructure is created for policies and strategic decisions using analyzed data.

Big Data Advanced Analytics Project – Web Scraping Method

Web Scraping Techniques:

- Job listings and price data are scraped daily using Python spiders.

Big Data Integration:

- The collected data is transferred to and managed in a big data environment via the Streamsets platform.
- Data flow is provided in real time with Apache Kafka.
- Data stored on HDFS can be analyzed with Impala.

Big Data Advanced Analytics Project – Web Scraping Result

- Within the scope of the project, daily job advertisements, agricultural products, and price data are systematically collected and analyzed using **web scraping** and **big data technologies.**

- The data obtained provides a reliable basis for statistical analysis and reporting processes.

- **Through to the project:**
  - Data collection and processing processes are automated,
  - Data quality is improved,
  - Contributions are made to statistical decision support mechanisms.

TURKISH STATISTICAL INSTITUTE

Thanks...