# Using Machine Learning Approaches for Economic Classification Based on Arabic Textual Descriptions

**Jaffar Mansour**
**RealSoft Company**
**jaffar.mansour@realsoft-me.com**

**Fatima Al Taharwah**
**RealSoft Company**
**fatima.altaharwah@realsoft-me.com**

## Abstract

Text data is one of the main pains of statistical survey analysis. The open-ended questions allow for obtaining data that is less biased by preconceptions. However, it takes a considerable amount of time and effort to interpret and analyze; thus, classification is usually performed by experts using standard classification systems like the International Standard Industrial Classification (ISIC) Rev. 4. Therefore, the time necessary to classify such data depends on the number of experts, this leads to an increase in effort and time as the volume of data increases.

To meet this challenge, Artificial Intelligence and data science can help research and statistics centers through automated classification of textual data using natural language processing techniques and text classification methods. The Arabic language constitutes a special topic in this study, as the data collected in statistical research in our region are usually in Arabic. Therefore, the required solution must consider the capabilities of the Arabic language and the challenges of its use.

One of the most important textual data collected by statistical centers is data on the economic activity of establishments and individuals. This data is usually collected in population censuses, economic surveys, and several social and economic studies. Benefiting from the information of the economic activity requires that it be classified according to ISIC Rev 4. Determining the correct classification for each unit would require efforts from specialists in industrial classification, and it is usually difficult to provide a sufficient number of these experiences in the time available to complete the statistical study and disseminate the results; moreover, a large number of statistical units can lead to human errors.

The main goal of this paper is to provide a review to find out the accuracy of the data cleaning and normalization, and classification technique in the automatic classification of economic activity textual data that are reported in Arabic.

Keywords: Text classification, Pre-processing, Text mining, Machine Learning, NLP, Arabic Language, ISIC, statistical survey analysis.

# 1. Introduction

Coding Economic Activity according to international and local standards is a critical yet demanding task faced by statistical institutions worldwide. This process consumes significant time and resources from employees, making it a prevalent international issue. However, with the help of machine learning, it's possible to build a computerized model that can learn from data and automatically classify new data into the correct categories.

When dealing with Arabic language data, specific challenges arise due to the intricate and multifaceted nature of the language. The Arabic language differs from other languages because of its complex and ambiguous structure that the computational system has to deal with at each linguistic level" [1]. Additionally, the processing of Arabic text lacks the necessary tools, resources, and algorithms from computational linguists, which results in its classification as a low-resource language within the natural language processing domain.

This paper aims to apply text classification for Economic Activity classification using real datasets from different countries. The main goal is to build a machine learning model that automates coding tasks for Economic Activity as international standard industrial classification (ISIC) to follow methodologies for International Labour Organization (ILO) and The United Nations Statistics Division (UNSTATS), which will enhance the time spent in this process and be more accurate than manual classification. The goal of this paper is achieved by building different machine learning models such as (Logistic Regression, Multinomial NB, Support Vector Machine, SGD Classifier, and Decision Tree Classifier), and comparing the results to recommend the best model to classify the Economic Activity. It also applies different experiments on the real data to improve models' performance. All experiments and models are evaluated using different evaluation measures such as accuracy, precision, recall, and F1 measure.

The idea of this paper was born in order to benefit from the great evolution of technology and artificial intelligence Besides that, most vectors tend to go through digital transformation projects; In addition, we aim to include our Arabic language in an important project that affects the statistics sector, which will enrich and increase the number of projects executed in Arabic, which is few now although Arabic is spoken by 280 million people and considered as the fifth used language in the world.

Arabic natural language processing (NLP) is an advanced application of AI, machine learning used to understand Arabic dialects; Arabic is considered a complex language syntactically and semantically because the Arabic language is used in 22 countries, people speak different accents and dialects even in the same city, which means an infinity number of words, vocabulary, and other details; also the plural in Arabic is mostly irregular form, and the same word has many synonyms.[1]

In addition, Arabic is divided into three main types, first is classic Arabic; the language of the Quran and ancient Arabic literature; the second type is the Modern Standard Arabic (MSA) which is simpler, is mainly used in formal places and situations like media, news, and official speeches, the third type which increase the difficulty is the spoken Arabic that is used in everyday life.

Many challenges face the Arabic language in NLP applications, like the lack of tools and resources, the lack of availability of a large collection in the few numbers of existing Arabic text classifiers; also the shortage of Arabic computational linguists and the combined expertise of NLP and Arabic Language is still not so common which normally increases the hindrances that data scientists face with NLP and indicates that more efforts have to be made in order to guarantee positive outcomes.[2]

In this paper, the researcher has faced many challenges, such as the lack of programming libraries and tools for Arabic, the familiar answer of a similar question must be short, which make it harder to encode , and multi-class problem, we have many categories around 400 for economic activities and this considered as a huge number for the text classification task, in addition to data bias, as the distribution of data tends to economic activities, so some activities contain a large number of data, on the other hand, the data is significantly few in other activities, this is an understandable bias that depends on the nature of countries and what they are known to manufacture, trade, or plant; while they have lack of dealing with other activities.

This paper is organized as follows: Section two presents ISIC as an international standard code for economic activity. In section three, we point out the Process Methodology, then Section four describes the dataset and the exploratory analysis of its features then we move to more specific details about the data preparation pipeline in Section five. In Section six, we discuss the experiments and result. Finally, Section seven draws our conclusion and future work.

## 2. About ISIC

The ICIS is a widely recognized standard classification system used to organize and classify economic activity data of establishments. This classification system categorizes institutions based on the nature of the activities they engage in. In most countries, the ISIC categories are determined by grouping similar activities described in statistical units, taking into consideration the significance of the activities included in each category. The current and most up-to-date version of ICIS is ICIS Rev.4 as shown in table 1 below.

Table 1: Broad sector concordance with ISIC

| **Aggregate Economic Activity** | | | **Sections ISIC-Rev. 4** |
|---|---|---|---|
| Agriculture | | | A |
| Non-Agriculture | Industry | Manufacturing | C |
| | | Construction | F |
| | | Mining and quarrying; Electricity, gas and water supply | B, D, E |

| **Aggregate Economic Activity** | | | **Sections ISIC-Rev. 4** |
|---|---|---|---|
| | Services | Market Services (Trade; Transportation; Accommodation and food; and Business and administrative services) | G, H, I, J, K, L, M, N |
| | | Non-market services (Public administration; Community, Social and other services and activities) | O, P, Q, R, S, T, U |
| Not elsewhere classified | | | X |

ISIC proved its effectiveness in many countries locally and internationally when it is used to classify data based on economic statistics studies like income, employment, and population.

Classification is important for giving substantial information to evaluate and observe economic performance and is also used to provide data for specific economic activities, as well as its importance in governmental uses.[3][4]

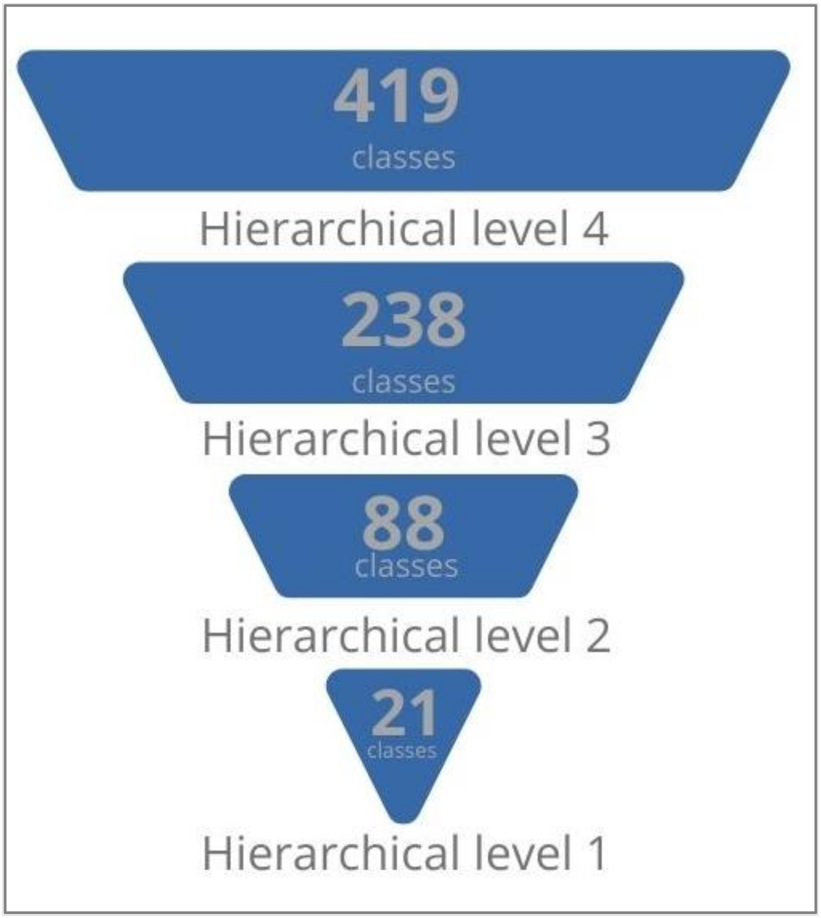

Figure 1:Annotation Hierarchical for ISIC4

## 3. Process Methodology

The implementation of this project relied on the Cross-Industry Standard Process for Data Mining (CRISP-DM), a well-established methodology that consists of six main phases aimed at making the process organized and efficient. Figure 2 outlines all the phases of CRISP-DM, starting from the crucial stage of business understanding, which guides the process and sets the objectives based on the gathered requirements and definitions. The following stage, data understanding, delves into the data's structure, nature, and issues to gain a deeper insight into it. Next, the data preparation phase focuses on cleaning and processing the data to bring it to an appropriate format that serves the desired purpose. Then, in the modeling stage, experiments are carried out, and different models and techniques are applied. In the evaluation phase, the work is estimated, and the best model and parameters are selected. Finally, the deployment phase is reached, where the golden model is developed. Throughout the successive stages of this research, all these phases were implemented with care and attention.

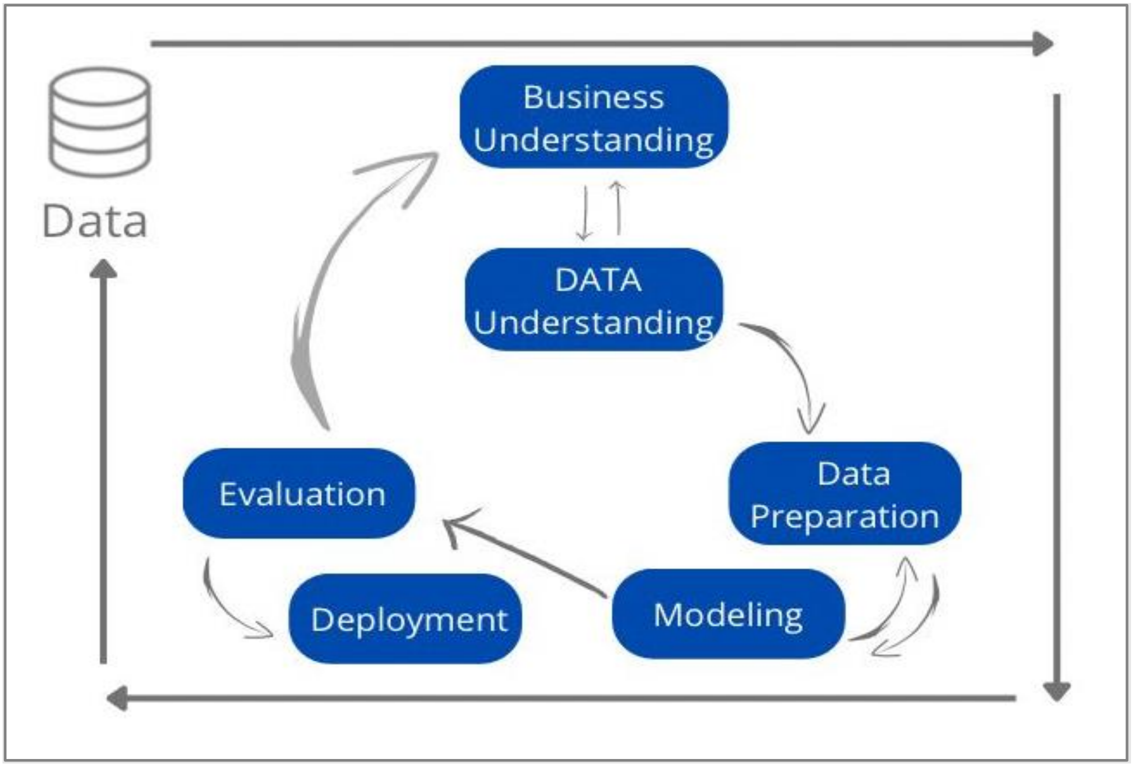

Figure 2:CRISP-DM Process Diagram

## 4. Data Source and Data Exploration

The dataset comprises multiple samples collected from different countries as part of regular surveys that gather various data points, including variables related to economic activities. One relatively large dataset was sourced from the Palestinian Central Bureau of Statistics' establishment census data, while the other datasets were obtained from economic survey data. Specifically, all data from the Department of Statistics and Community Development in Sharjah, the Ras Al Khaimah Statistics Center, and the National Institute of Statistics in the Republic of Tunisia were based on economic surveys. After cleaning, all four datasets were split into training and test sets.

The combined dataset consisted of a relatively large collection of raw descriptions of Economic Activities in Arabic, as defined by surveyors, along with their corresponding codes from the fourth level of the ISIC4 standard (supervised data). The coding was carried out by experts from the owning statistical office. The dataset comprised over 180,000 records and included more than 400 categories. As the dataset did not include any identifiable personal or entity information, but only text for economic activity description and coded fields, we confirmed that there were no ethical considerations in using it.

By applying Exploratory Data Analysis (EDA) techniques, we were able to gain valuable insights and information about our data source. Through EDA, we identified patterns and relationships within the data, evaluated its quality, and pinpointed issues that needed cleansing and transformation.

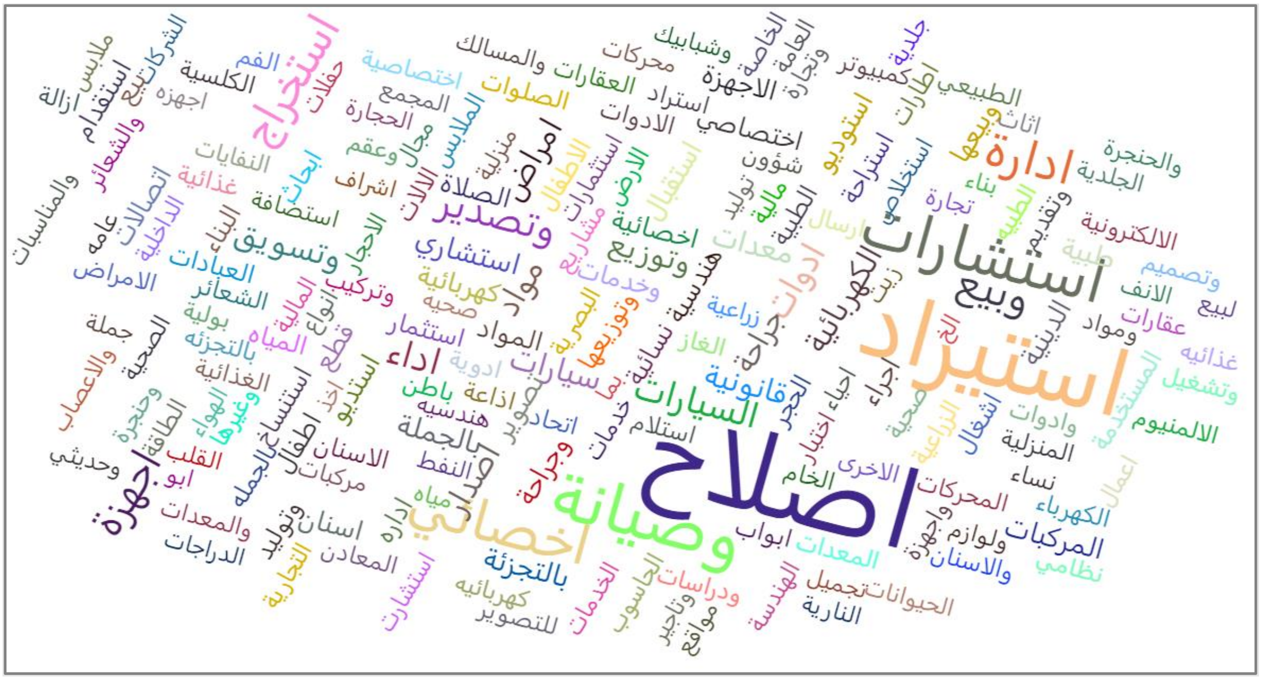

Figure 3: Word Cloud for Corpus

## 5. Data Preparation

The data preparation phase is crucial in any study as it can significantly impact the subsequent phases and final outcomes of the research. In natural language model building, this phase requires specific preprocessing steps to ensure that the text data is cleaned and optimized for analysis. Therefore, the researchers placed significant value on data preparation and built an outstanding preprocessing pipeline containing several functions to handle various issues. Some of these functions aimed to deal with general noise in the data, such as white spaces, hashtags, numbers, links, punctuation marks, and special characters. Meanwhile, other functions addressed the specific characteristics of the Arabic language, such as "Tatweel," "Tashkeel," "Alhamzah," and Arabic stop words.[5]

To ensure a thorough and effective data processing, the researchers divided the process into five essential stages, as shown in Figure 4: Normal Cleaning, Noise Removal, Normalization, Tokenization, and Vectorization using Term Frequency-Inverse Document Frequency (TF-IDF). The TF-IDF technique is a statistical method for evaluating the importance of a word based on its frequency of occurrence in the document and its relevant corpus. By implementing these data preparation techniques, we could optimize the data for analysis and maximize the accuracy and reliability of the final results. [6][7]

Table 2: TF-IDF Formula

| Mathematical forms of TF-IDF = (tf-idf (t, d) = tf (t, d) * log (N/ (df + 1)))<br>Where: t = term (word), d = document (set of words), N = count of corpus, corpus = the total document set | |
|---|---|
| Document Frequency | df(t) = occurrence of t in documents |
| Term Frequency | tf (t, d) = count of t in d / number of words in d |
| Inversed Document Frequency | idf(t) = log (N/ (df + 1)) |

The preprocessing process is complicated when the texts are short [8], and each phase of them contains multiple tasks that help to make the quality of data better, reduce the size of the input data, package the data for modeling, and minimize this overhead in the next phases.

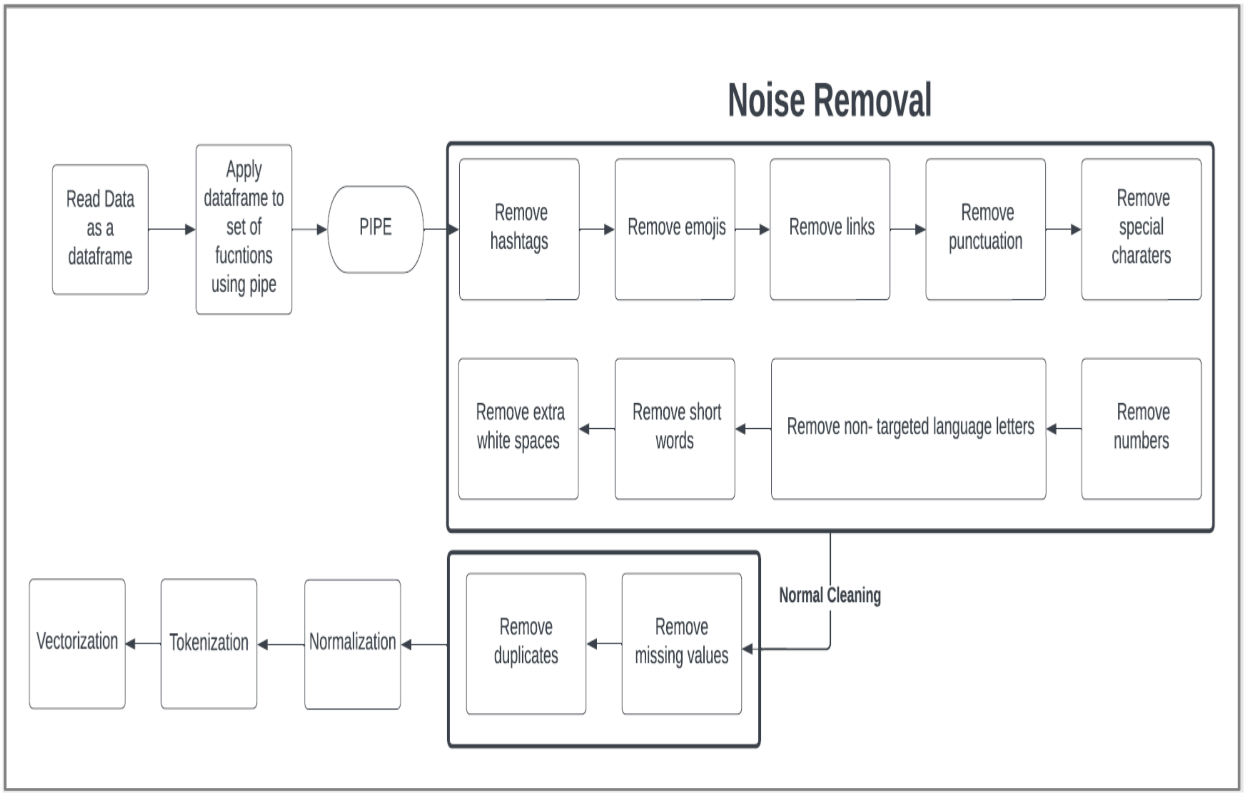

Figure 4:Preprocessing Pipeline

## 6. Experiments and Results

The conventional approach to text classification typically involves several fundamental stages that encompass data pre-processing, feature extraction, selection of a suitable machine learning model for classification, training the classifier, and ultimately, testing the classifier using the trained model.

Following the completion of pre-processing techniques, the researchers conducted experiments on the combined dataset. The dataset was thus rendered ready for modeling and the application of machine learning algorithms.

The data was partitioned into two sets: input (X), which comprises the raw descriptions of economic activities, and output (Y), which represents the class label of ISIC4 codes. Additionally, we randomly divided the dataset into training and testing subsets using different percentages in our experiments, such as 90/10, 80/20, and 70/30.

Various techniques have been proposed for automatic classification of literature data, but there is no consensus on the best approach, as the choice of method depends on the specific characteristics of the data being analyzed. Therefore, we have opted to implement a range of algorithms and conduct a comparative analysis between them. Some of the algorithms we have considered include:

Logistic Regression: It is a probabilistic based on a statistical model used to solve classification issues in machine learning. Logistic regression typically uses a logistic function to estimate the probabilities. [9]

- Multinomial NB: It is a model that assumes that all attributes are independent of each other given the context of the class; it ignores all dependencies among attributes. [9] it is a probabilistic learning method that is mostly used in Natural Language Processing (NLP). The algorithm is based on the Bayes theorem and predicts the class of a text, such as a piece of email, or newspaper article. It calculates the probability of each class for a given sample and then gives the class with the highest probability as output.

- Stochastic gradient descent (SGD): is an iterative method for optimizing an objective function with appropriate smoothness properties. [9]

- Decision Tree Classifier: It works by creating a tree-like model of decisions and their possible consequences, based on the features of the input data, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions. [10] [11]

Various methods are conventionally used to evaluate the effectiveness of a model, including precision (true positives / predicted positives), recall (true positives / all actual positives), F1 (2 * (precision * recall) / (precision + recall)) and accuracy (all correct / all). However, accuracy is not a recommended metric for unbalanced datasets and can be biased in such cases [12]. Therefore, the classification performance of the algorithms in this paper is measured by precision, recall, and the F1 score, which is a combination of both precision and recall. The researchers conducted multiple experiments to achieve their goal, but we will focus on discussing three of them. The first experiment starts with applying the four algorithms mentioned above on the cleaned data and the representation of TF-IDF which uses term weight to enhance text data representation as the feature vector.

The empirical results show the acceptable result, as shown in table 3 below, which clarifies fitting the model using the training data and evaluating the performance.

Table 3:Final Results Before Handling Imbalanced Data

| Model Name | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Logistic Regression | 0.78 | 0.80 | 0.78 | 0.80 |
| Multinomial NB | 0.67 | 0.64 | 0.67 | 0.60 |
| SGD Classifier | 0.79 | 0.77 | 0.79 | 0.76 |
| Decision Tree Classifier | 0.75 | 0.75 | 0.75 | 0.75 |

Then in the second experiment, the researcher tried to enhance the results by reducing the gap between the different classes (minority classes and majority classes) to solve the unbalancing problem using approaches such as Remove the class with less than 10 examples from the dataset, also oversampling to increase the sample for minority classes using SMOTE (Synthetic Minority Over-sampling Technique) which is an approach handle the inequality between the categories representation.[13]

Table 4:Final Results After Handling Imbalanced Data

| **Model Name** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
|---|---|---|---|---|
| Logistic Regression | 0.96 | 0.96 | 0.96 | 0.95 |
| Multinomial NB | 0.92 | 0.93 | 0.92 | 0.92 |
| SGD Classifier | 0.86 | 083 | 0.86 | 0.85 |
| Decision Tree Classifier | 0.98 | 0.98 | 0.98 | 0.98 |

We can see the result enhancement after using oversampling. Table 4 shows the result of applying an innovative approach to all the data sets, with results above 80% in all algorithms.

The third experiment performs an interesting technique is Binary classification to multiclass classification One-vs-Rest (OvR)

In general, using algorithms like SVM and logistic regression in binary classification studies can help to determine one of two classes; therefore, classification is considered as binary classification case.

However, there are many cases where have more than two classes, so we need an algorithm that is able to work with multiple classes.

There are many available samples for this type of classification, and there are many techniques that can make support vector machines able to dissolve multiple classes; one of them is One-vs-Rest, as figure 5 shows.[14][15]

This experiment took a long-time during model training stage, and this was understandable because the technique attempted to build 419 models, which is the total number of economic activities at the fourth level, as it transforms each economic activity into a binary model that represents the activity against the rest of the other categories. Unfortunately, despite the time and effort spent in this experiment, the results were inconsistent. Satisfactory with an accuracy of only 71%, and this may be acceptable for the massive number of categories in this data.

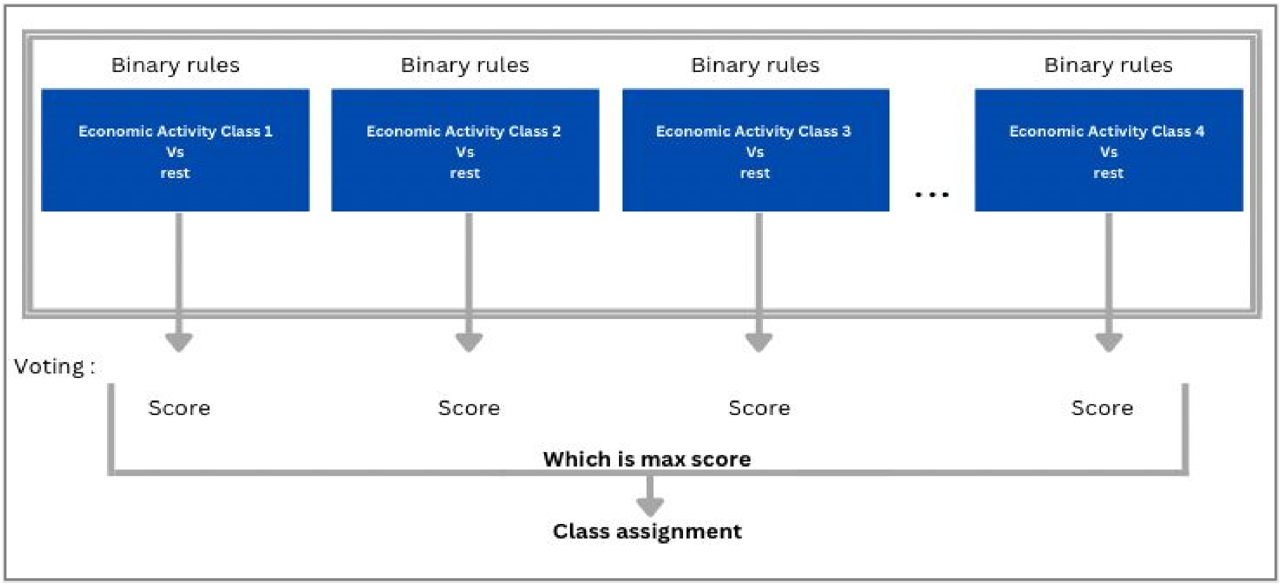

Figure 5 : One-vs-rest Classifier

## 7. Conclusion and Future Work

This paper presents the collection of a raw Economic Activity Data set in Arabic, along with its corresponding ISIC 4 classification, from four National Statistical Organizations. The raw data has been consolidated into a unified dataset and has undergone successful analysis, transformation, and visualization. Additionally, a pre-processing technique has been developed to effectively address generic text quality issues as well as Arabic text-specific qualities. This has resulted in an exceptional tool for handling Arabic Economic Activity Data. Five machine learning models were used to predict international standard industrial classification (ISIC) code with a lot of effort and many iterations to improve the results. Convergent and satisfactory results exceeded 90 percent for most of the algorithms, and we are still trying to test the models on real data to ensure the efficiency of the evaluation.

It has been noticed that even for the same classification way, classifier work can differ for each training text corpuses, and some differences can be substantial. This observation proves that classifier performance is partly based on the input training corpus.

Our methods have demonstrated statistical significance in successfully tackling various challenges relevant to imbalanced dataset problems, the intricate nature of Arabic text processing, and the management of a high number of classes. Incorporating machine learning (ML) technologies could offer substantial time and cost savings for official statistics agencies in Arabic-speaking countries when it comes to categorizing unprocessed economic activity descriptions into International Standard Industrial Classification (ISIC) 4 coding systems for the majority of Economic Activities (EA) classes. Our future research would also include the utilization of additional deep and machine learning models. Furthermore, we plan to expand the size of our training dataset by incorporating as much data as possible from various surveys and Arab countries. By doing so, we can enhance the coverage of economic activities classes and increase the diversity of dialects across multiple Arab countries. In our future research, we aim to discover more effective techniques and tools to address additional aspects of the Arabic language that were not addressed in this study. We will investigate techniques for handling hierarchical structures of Classes, and, importantly, we will concentrate on identifying more robust methods for managing the immense number of classes that we encounter. We believe that our research will make a valuable contribution to statistical studies conducted by official statistics offices by reducing the cost and time required to encode economic activities. Furthermore, we hope that our findings can be applied to alleviate the human-intensive work involved in classifying other statistical variables, thus enabling more extensive research in this field.

## References

[1] Abdeen, M. A., AlBouq, S., Elmahalawy, A., & Shehata, S. (2019). A closer look at arabic text classification. International Journal of Advanced Computer Science and Applications, 10(11).

[2] Elayeb, B. (2021, November). Arabic Text Classification: A Literature Review. In 2021 IEEE/ACS 18th International Conference on Computer Systems and Applications (AICCSA) (pp. 1-8). IEEE.

[3] https://unstats.un.org/unsd/classifications/Family/Detail/27

[4] International Standard Industrial Classification of All Economic Activities manual.

[5] Said, D., Wanas, N. M., Darwish, N. M., & Hegazy, N. (2009, April). A study of text preprocessing tools for arabic text categorization. In The second international conference on Arabic language (pp. 230-236).

[6] Zhang, W., Yoshida, T., & Tang, X. (2008, October). TFIDF, LSI and multi-word in information retrieval and text categorization. In *2008 IEEE International Conference on Systems, Man and Cybernetics* (pp. 108-113). IEEE.

[7] Jones, K. S. (2004). IDF term weighting and IR research lessons. *Journal of documentation.*

[8] Bobicev, V., & Sokolova, M. (2008, July). An Effective and Robust Method for Short Text Classification. In AAAI (pp. 1444-1445).

[9] Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. SN Computer Science, 2(3), 1-21.

[10] Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

[11] Saad, M. K., & Ashour, W. M. (2010). Arabic text classification using decision trees. In Proceedings of the 12th international workshop on computer science and information technologies CSIT (Vol. 2).

[12] Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text classification using machine learning techniques. WSEAS transactions on computers, 4(8), 966-974.

[13] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

[14] Hong, J. H., & Cho, S. B. (2008). A probabilistic multi-class strategy of one-vs.-rest support vector machines for cancer classification. Neurocomputing, 71(16-18), 3275-3281.

[15] Faris, H., Habib, M., Faris, M., Alomari, M., & Alomari, A. (2020). Medical speciality classification system based on binary particle swarms and ensemble of one vs. rest support vector machines. Journal of biomedical informatics, 109, 103525.