The Scientific Journal of Cihan University – Sulaimaniya **PP: 140-157**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Kurdish Sorani Natural Language Processing Text Classification Using Different Machine Learning Algorithms

Orme Mhamad Amin[1], Nzar Abdulqadir Ali[2]

[1]Computer Institute of Sulaimani, Sulaimani, Iraq
[2] Department of Statistics and Informatics, University of Sulaimani, Sulaimani, Iraq

Email: urme.mhamad85@gmail.com[1], nzar.ali@univsul.edu.iq[2]

Abstract:

Researchers assessed machine learning models on Kurdish Sorani natural language progression classification utilizing dataset information obtained from different Kurdish news sites. The collected dataset contains five categories which are business, sport, health, social and technology. The research tested Extreme Gradient Boosting (XGBoost) as well as Random Forest (RF) alongside Support Vector Machine (SVM) algorithms and Light Gradient Boosting Machine (LGBM). The results show that (SVM) prove the most effective performance reaching the maximum accuracy and F1-scores across all classification categories. Although XGBoost displayed good performance in classifying technology and culture news Random Forest and LGBM struggled to realize accurate outcomes for the social class category. The study highlights the potential of modern machine learning models for attractive text classification tasks.

Keywords: News Classification; Natural Language Processing; Machine Learning.

الملخص:

قيّم الباحثون نماذج التعلم الآلي في تصنيف اللغة الكردية السورانية باستخدام مجموعة بيانات تم جمعها من عدة مواقع إخبارية كردية. تحتوي هذه المجموعة على خمس فئات وهي: الأعمال، الرياضة، الصحة، المجتمع، والتكنولوجيا. تم اختبار نماذج Extreme Gradient boosting (XGBoost) , Random Forest (RF) , Support Vector Machine (SVM) , Light Gradient Boosting Machine (LGBM). أظهرت النتائج أن(SVM)كانت الأكثر فعالية، حيث حققت أعلى دقة و F1-scoreفي جميع فئات التصنيف. وعلى الرغم من أن نموذج XGBoost أظهر أداءً جيدًا في تصنيف الأخبار التكنولوجية والثقافية، إلا أن نماذج Random Forest و LGBMواجهت صعوبة في تحقيق نتائج دقيقة لفئة الأخبار الاجتماعية. تسلط الدراسة الضوء على إمكانات نماذج التعلم الآلي الحديثة في مهام تصنيف النصوص كردية.

الكلمات المفتاحية: تصنيف الأخبار؛ معالجة اللغة الطبيعية؛ التعلم الآلي.

پوختە:

لەگەڵ بەکارهێنانی زانیاریەکان کە لە ماڵپەڕەکانی هەواڵی کوردی وەرگیراون، توێژەران مۆدێلی فێربوونی ماشینیان هەڵسەنگاند بۆ ڕیزبەندی زمانی سۆرانی کوردی. داتاسێتەکە پێکهاتووە لە پێنج بەش لەوانە بازرگانی، وەرزش، تەندروستی، کۆمەڵایەتی و تەکنەلۆجیا. لە توێژینەوەکەدا چوار مۆدێلی تاقیکراونەتەوە لەوانە Extreme Gradient boosting (XGBoost) , Random Forest (RF) , Support Vector Machine (SVM) , Light Gradient Boosting Machine (LGBM) ئەنجامەکان دەریدەخەن کە (SVM) باشترین دەرەنجامی نیشانداوەو بەرزترین ڕێژەی دروستی و F1-score لە هەموو بەشەکاندا. هەرچەندە XGBoost کاریگەرییەکی باشی هەبوو لە بەشەکانی تەکنەلۆجیا و کەلتوردا بەڵام Random Forest و LGBM

DOI: http://dx.doi.org/10.25098/9.1.32

The Scientific Journal of Cihan University – Sulaimaniya **PP: 140-157**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

نەیانتوانی پۆلێنی بەشی کۆمەڵایەتی بە باشی بکەن. ئەم توێژینەوەیە گرنگی بەکار هێنانی مودێلەکانی (ML) دەردەخات بۆ جیاکاری نوسینەکان لە زمانی کوردی دا.

کلیلە وشە: پۆلێنکردنی هەواڵ؛ پرۆسێسی زمانی سروشتی؛ فێربوونی ئامێر.

1.Introduction

Since developing technology so fast the internet has become an important component of human daily life, many data type is available on the internet, text is one of the most common forms. Kurdish language is one of the low-resource language, around 40 million speakers worldwide, mainly in Iran, Syria, Turkey and Iraq. Central Kurdish Language (CKL), also known as Sorani, represents one of Kurdish dialect within the broader Kurdish linguistic landscape. [1]

Machine learning (ML), a subset of artificial intelligence (AI), involves enabling machines to learn from historical data and improve their accuracy over time without explicit programming. [275] In recent years, the integration of machine learning techniques into the field of natural language processing (NLP) has transformed various paradigms of textual analysis [1,2]. It is difficult for human to collect those amounts of data manually for this reasons Machine learning techniques can be applied to the data that the dataset can be extracted from it, it will be able to help us in many fields such as get information about how we should able to use available resources and making a decision about the future [2].

Text Classification (TC) stands as a critical task, important to NLP, ML and the realm of Data Mining (DM). In fact, TC has grown into a method for extracting, detecting, analyzing, and retrieving user knowledge from large bodies of text. Broadly, TC can be divided into two main types: multiclass and multi-label classification. In multiclass classification a single piece of text can be tagged with one label, in time in multi-label a piece of text tagged with more than one label such as “health” and “social”.

It is an important to note that the accuracy of NLP models depending on the quantity and quality of data used for training. However, it is more difficult for collecting enough labelled data for under-resource language need more time, funding and resources [3].

The limited availability of annotated datasets and linguistic tools specific to the Kurdish language has created barriers for implementing effective machine learning algorithms for tasks such as text classification [4]

Text categorization usually involves a number of different ML algorithms. These include, but are not limited to, Support Vector Machines (SVMs), Random Forest (RF), Naive Bayes classifiers, Decision Trees, K-Nearest Neighbor, Light Gradient Boosting Machine (LightGBM), Extreme Boosting (XGBoost) [5]

DOI: http://dx.doi.org/10.25098/9.1.32

The Scientific Journal of Cihan University – Sulaimaniya **PP: 140-157**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

2. Literature Review

(Abdullah, Muhamad, et al., 2024) creating efficient text-to-speech (TTS) for low-resource languages such as Central Kurdish (CKB) are addressed in this work are very difficult. A Kurdish WaveGlow vocoder using a specific 21-hour CKB speech corpus, rather than using pre-trained English models, the study aimed to greatly enhance Kurdish TTS by training. With an astounding Mean Opinion Score (MOS) of 4.91 the adaptive WaveGlow model's output established a new standard for Kurdish voice synthesis and demonstrated the vital necessity of language-specific vocoder training for fluent and genuine speech. The major result is a Kurdish TTS system that is much improved. This work not only enhances Kurdish TTS but also provides a scalable approach for additional low-resource languages. [6]

(Badawi et al. 2024) for Kurdish news article classification Introduces RFO-CNN a revolutionary hybrid method. The suggested approach fine-tunes the CNN's parameters by joining a convolutional neural network (CNN) with an enhanced version of the red fox optimization algorithm (RFO).The KNDH and KDC-4007,are two popular Kurdish news that using the data set the authors evaluated the RFO-CNN model and contrasted its results with those of other deep learning models such as BERT and BLSTM and traditional machine learning techniques. Depending on the results the RFO-CNN model performs better in terms of accuracy and F1-score in a variety of training and testing settings than the benchmark BERT model and other machine learning models. According to the authors the RFO-CNN model offers intriguing opportunities for optimizing deep learning architectures for underrepresented languages and domains. They also stress the significance of using customized algorithms for particular low-resourced languages. [7]

(Ahmad & Rashid, 2024c) Although single-stage processing is the goal of recent developments in text-to-speech (TTS) models, audio quality is frequently subpar, especially for Kurdish. This study begins by emphasizing the urgent need for better TTS for the Kurdish language, given the underrepresentation of Kurdish, particularly the Sorani dialect, in these developments. The result of this study is to present KTTS an effective end-to-end TTS model for creating high-quality Kurdish audio. A variational autoencoder (VAE) that has been pre-trained for audio waveform reconstruction is used to accomplish this along with adversarial training to align latent distributions and a stochastic duration predictor for a variety of speech rhythms. The main result is the creation and practical testing of KTTS on a customized dataset which shows subjective human evaluation of its improved performance (MOS of 3.94) in comparison to one-stage and two-stage baseline systems. Future work is recommended to improve the model, broaden its application to various Kurdish dialects and styles, and integrate the VAE and KTTS training procedures. This study provides high-quality, an effective, and adaptable strategy for Kurdish TTS.[8]

(Rawf et al. (2024) This paper first highlights the current dearth of appropriate datasets for developing automatic dialect recognition systems specifically for the Kurdish language, acknowledging that speaker variations such as dialect have a significant impact on speech recognition system performance and that integrating a Dialect Recognition System (DRS) can mitigate this by choosing the appropriate speech recognition model. By evaluating a recently presented dataset that was gathered by academics at the University of Halabja's Computer Science Department and

DOI: http://dx.doi.org/10.25098/9.1.32

The Scientific Journal of Cihan University – Sulaimaniya **PP: 140-157**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

comprises the three primary Kurdish dialects—Northern Kurdish (Badini), Central Kurdish (Sorani), and Hawrami—this effort aims to close this gap. The offered material concentrates on the rationale and the dataset itself, without specifically stating the outcome of the dataset evaluation or the performance of any DRS constructed using it. Consequently, it is impossible to establish a definitive "result" based only on this text. By presenting a new resource, the report lays the groundwork for further research in Kurdish dialect identification.[9]

(Saeed, 2024b) using four Kurdish text datasets Kurdish News Dataset Headlines (KNDH), Medical Kurdish Dataset (MKD), Kurdish-Emotional-Dataset (KMD-77000) and KurdiSent. This paper presents FastText, a novel text classifier for the Kurdish language, and evaluates its performance against both classic machine learning (Random Forest, k-NN, Logistic Regression, Multinomial Naïve Bayes, Support Vector Machine, Decision Tree, Stochastic Gradient Descent) and a deep learning model (BERT).Across all datasets, the study discovered that FastText consistently performed better than any other model. In specific Fast Text outperformed KurdiSent (81.32% precision, 81.83% recall, 81.57% F1-score, and 81.4% accuracy), MKD (93.32% precision, 93.36% recall, 93.34% F1-score, and 93.1% accuracy), KMD (about 2% better than others), and KNDH (89% precision, recall, F1-score, and 89.10% accuracy). Before dividing the data into training and testing sets the researchers preprocessed it using stop word removal, lemmatization, tokenization, and stemming. The discoveries unequivocally show that FastText is the best classifier for classifying Kurdish texts. In order to further improve text categorization in Kurdish the paper proposes that future research investigate hybrid models based on FastText.[10]

(Shareef et al. (2024) this study proposes a deep learning model based on Mask R-CNN for digitizing Kurdish handwritten text recognition (KHTR) In order to meet the demand for technology to transform paper-based records in the Kurdistan Region of Iraq into digital text for e-government services, To test their methodology, the researchers produced a dataset of 40,410 isolated handwritten Central Kurdish character pictures from 390 native writers. The suggested model outperformed on a difficult dataset comparable study with over 99% OCR accuracy, achieving promising results with 80% precision, 96% recall, and an 87.6% F-score. Using a newly constructed Local Challenges Text (LCT) dataset, the study addresses problems such as different font types, sizes, colors, orientations, and image illumination, highlighting the originality of employing Mask R-CNN for simultaneous text detection and localization. The procedure includes word separation, character segmentation, background removal using bounding box and segmentation, multi-level feature extraction using MSER on Mask R-CNN results, and the potential use of OCR on video frames. In addition to highlighting how their model can help researchers and the Kurdish community by enabling more accurate and efficient digitization of historical archives and handwritten documents, the authors draw the conclusion that deep learning can significantly improve e-government services and preserve cultural heritage. Future work will focus on building larger datasets and exploring other deep learning models in order to significantly increase recognition accuracy. [11]

(Ahmad et al. (2024) this study explains the problems for Central Kurdish in creating effective speech synthesis systems which are mostly caused by modeling problems and a lack of effective training and analysis techniques. Even though there are about 30 million Kurdish speakers globally the lack of text and speech recognition capabilities is the biggest obstacle to Kurdish text-to-speech

DOI: http://dx.doi.org/10.25098/9.1.32

The Scientific Journal of Cihan University – Sulaimaniya **PP: 140-157**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

(TTS). In order to address this the paper, explain the Kurdish Text-to-Speech Dataset (KTTS, Gigant), a huge vocabulary dataset that contains a speech corpus and a pronunciation lexicon for the Central Kurdish dialect. In order to faithfully capture real-world Central Kurdish, a male Kurdish dubber recorded 6,078 phrases over 12 document topics in a professional studio. This voice corpus aims to serve as the main source of information for creating efficient speech systems for this language. [12]

(Abdullah, Abdulla, et al., 2024b) Due to the language's diverse structure, scarcity of datasets, diversity of dialects, and Kurdish Natural Language Processing (KNLP) presents several difficulties, which are especially evident in Kurdish Named Entity Recognition (KNER). In order to overcome this the researchers, suggest optimizing a pre-trained RoBERTa model for KNER. They developed a modified model architecture put training methods into place, and produced a Kurdish corpus. Setting a new standard for KNLP, the experimental results demonstrated that fine-tuning RoBERTa with sentence-piece tokenization greatly enhanced KNER performance by 12.8% in F1-score when compared to traditional models. The emphasized in the paper is significance of sophisticated transformer architectures and fine-tuning for low-resource languages. While recognizing the continued difficulties of sparse annotated data, intricate Kurdish morphology and dialect variances, future research will concentrate on growing datasets and adding new dialects to improve model resilience. Zero-initialized attention is also examined in the research as a computationally effective adaptation technique. [13]

(Mahmud et al., 2023) striking lack of attention on Kurdish Sorani in sentiment analysis research within Natural Language Processing in contrast to the significant quantity of work done for English is the examines of this paper. By creating a new Kurdish sentiment analysis dataset for social media language and assessing how well several machine learning (Random Forest, KNN, SVM, Naive Bayes, Decision Trees) and deep learning (ANN, LSTM, CNN) approaches performed on this dataset the study wanted to bridge this gap. The primary conclusion was that the Naive Bayes algorithm outperformed the others with an accuracy of 78%. The study highlighted the positive impacts of F-5 stemming on the results for all models even as it noted that the performance of deep learning models was inconsistent most likely due to the short dataset size. The authors recommend enlarging the dataset and looking into emotion analysis in their subsequent endeavor.[14]

(S. Badawi, 2023) This paper begins by pointing out the lack of resources for Kurdish text summarization, given the growing amount of digital information and the significance of text summarization in domains such as information retrieval and natural language processing. It also notes that current research focuses on languages like English and Chinese with little consideration for Kurdish. With over 40,000 manually summarized news articles, KurdSum is the first Kurdish summarization news dataset ever created. Its performance will be evaluated using a variety of extractive (LEXRANK, TEXTRANK, ORACLE, LEAD0-3) and abstractive (Pointer-Generator, Sequence-to-Sequence, transformer-abstractive) summarization techniques. The main outcome is the development and assessment of the KurdSum dataset, which shows that ORACLE performed better than extractive techniques while the Pointer-Generator abstractive approach obtained higher ROUGE ratings. KurdSum, according to the authors, provides a potential path for Kurdish text summarizing and is a useful standard for further study and the creation of NLP tools for the Kurdish language.[15]

DOI: http://dx.doi.org/10.25098/9.1.32

The Scientific Journal of Cihan University – Sulaimaniya **PP: 140-157**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

(Badawi et al., 2023)The important challenge of producing a high-quality dataset for Kurdish text categorization is addressed in this study. The Kurdish News Dataset Headlines (KNDH) a sizable resource with 50,000 news headlines evenly split over multiple categories (10,000 headlines each class) was the main goal of this study. The Pars Hub tool and the BeautifulSoup library were the primary web scraping techniques utilized to obtain these headlines from 34 distinct Kurdish news sources. As a result, even if the distribution of headlines from different news channels varies, the KNDH provides a balanced dataset with equal representation across categories. To guarantee a clean and usable resource for training and assessing text classification models for the Kurdish language, the gathered data underwent necessary preprocessing using the Kurdish Language Processing Toolkit (KLPT), including tokenization, spell-checking, stemming, and removal of irrelevant elements.[16]

3. Background Theory

3.1 TFIDF Vectorizer

TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical way of figuring out how important a word is in a particular document, especially when one compare it to all the other documents (corpus). It is a method for text processing that converting text into numbers so machines can learn from it. The idea is to distill words into number values. It generally happens in two major steps. [ww6]

1-Term Frequency essentially tells us how often a word occurs in a specific document. Generally, the more a word shows up in a given text, the more important it is, at least within that particular document.

$$TF(t,d) = \frac{Number\ of\ times\ term\ t\ apperas\ in\ document\ d}{Total\ number\ of\ terms\ in\ document\ d} \quad \text{.....(1)}$$

2- IDF works by looking at how often a word appears in a whole bunch of documents, not just one. Basically, if a word shows up everywhere, it's considered less important.

$$IDF(t) = \log\left(\frac{Total\ Number\ of\ Document}{Number\ of\ Documents\ containg\ term\ t}\right) \text{.....(2)}$$

TF-IDF are combining these two processes for giving the weight for each word in each topic, the word that have the highest scores are noted as important feature for the machine learning model, that allows the model to evaluate the relevance of words in each topic ultimately enabling it to effectively classify the text data.

DOI: http://dx.doi.org/10.25098/9.1.32

The Scientific Journal of Cihan University – Sulaimaniya **PP: 140-157**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

3.2 Machine Learning Algorithms

1- Random Forest (RF)

A Random Forest algorithm is a powerful and useful ensemble knowledge method generally working for each classification and regression mission in machine learning. It consisting of many decision trees that work organized for making guesses each tree is trained on a random subset of the data tested with replacement and considers only a random subset of features at each split.[17] This deliberate introduction of randomness ensures variety among the different trees. When presented with a new instance each tree independently generates a prediction. For classification problems the final prediction is determined by a majority vote across all trees while for regression It is the average of their predictions. This ensemble approach, where individual potentially weak learners collaborate significantly enhances prediction accuracy and robustness by mitigating overfitting. Also, random forest inherently handles missing data provide a measure of feature importance and scale effectively to large and complex data sets another they can be computationally exhaustive and less explainable than single decision trees. [18]

2-Support vector machine (SVM)

Is a strong supervised algorithms that used for classification aiming to identify the optimal hyperplane that separate data of different classes in an N-dimensional feature space. This is realized by maximizing the margin the distance between the hyperplane and the nearest data points known as support vectors. For linearly separable data SVM seeks a (hard margin) but for real world scenarios with potential overlaps a (soft margin) is working allowing some misclassifications measured by a regularization parameter (C) to improve generalization. When dealing with non-linearly separable data SVM uses kernel functions to indirectly map the data into a higher-dimensional space where linear separation converts feasible. The type of Common kernel contains: polynomial, linear, and Radial Basis Function (RBF). Though SVM excels in high-dimensional spaces handles non-linear relationships over kernels demonstrates resilience to outliers by soft margins and supports both binary and multiclass classification with memory efficiency it can suffer from slow training on large datasets needs careful parameter tuning (kernel and C) is sensitive to noise and feature scaling and offers limited interpretability due to the difficulty of the separating hyperplane in higher dimensions.[19]

2- Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting or XGBoost is a commonly used machine learning algorithm and highly effective that controls an ensemble of decision trees built serially. Unlike random forests that build trees in parallel XGBoost concepts each new tree to correct the errors made by the preceding trees thereby iteratively improving prediction accuracy and minimizing the loss function. The boosting practice joins advanced optimization methods and regularization to avoid overfitting and enhance overall model performance. Careful adjustment of the parameters is essential for optimizing XGBoost models for specific real world machine learning tasks.[20]

DOI: http://dx.doi.org/10.25098/9.1.32

The Scientific Journal of Cihan University – Sulaimaniya **PP: 140-157**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

3- Light gradient boosting machine(lightGBM)

Light GBM is an attractive option for a variety of supervised learning tasks such as classification, regression, and ranking, due to its speed, accuracy, memory efficiency, and capacity to manage huge datasets. It uses cutting-edge methods like as Gradient-based One-Side Sampling (GOSS), histogram-based algorithms, and leaf-wise tree development to build powerful learners by gradually adding weak decision trees in a gradient descent fashion. LightGBM frequently outperforms other boosting frameworks, these capabilities which also allow it to achieve quicker training speeds and reduced memory usage. [21]

3.3 Evaluation Metrics

Model performance in machine learning is critical. The right metrics can guide model selection, optimization, and deployment. The effectiveness of a binary classification model can be measured by confusion matrix (Figure1). Basically, it's a table offering insights into the model's accuracy. The table presents predicted values Positive (1) or Negative (0) alongside the real, "True" (1) or "False" (0) target values. Furthermore, from this matrix, we can figure out several metrics that reflect how well the model is performing. These include overall accuracy, and then more metrics like precision, recall, and F1 score.

- True Positive (TP): This arises when the model correctly predicts a positive outcome, and that outcome is, in fact, positive
- True Negative (TN): This arises when the model correctly predicts a negative outcome, and that outcome is, in fact, negative
- False Positives (FP): a scenario where the model suggests a positive outcome, but the reality is negative (Type I error)
- False Negative (FN): a scenario where the model suggests a negative outcome, but the reality is positive (Type I error)

Class designation		Actual class	
		True (1)	False (0)
Predicted class	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 1. Binary Classification Confusion Matrix [22]

The quantities TP, TN, FP, and FN, used to compute a range of performance indicators, as shown in Table (1)

DOI: http://dx.doi.org/10.25098/9.1.32

The Scientific Journal of Cihan University – Sulaimaniya **PP: 140-157**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Table (1). The Elements of the Evaluation Process

Variable	Definition	Equation
Accuracy	Accuracy measures the proportion of true results among the total cases examined	$Accuracy = \frac{Tp+TN}{TP+TN+FP+FN}$
Precision	Precision focuses on the accuracy of positive predictions	$Precision = \frac{TP}{TP+FP}$
Recall	Recall emphasizes a model's ability to identify all relevant instances	$Recall = \frac{TP}{TP+FN}$
F1-Score	The F1 score, harmonizes precision and recall into a singular metric, proves particularly useful when balancing these competing interests is critical	$F1 - Score = 2 \times \frac{percision \times recall}{Percison+recall}$

4. Methodology

This paper aims to classifying Kurdish sorani text into multiple categories like (sport, health, economy, business, social) by creating an evaluating machine learning model. The methodology contains five steps: data collection, preprocessing, handling imbalance data, model selection and evaluation. In the first stage the dataset must be loaded and restricted by maximum of 1000 samples per category to ensure a balanced representation across different categories,

TF-IDF (Term Frequency-Inverse Document Frequency) vectorization implemented to converts text into numerical data; it essentially picks out what's most important in the text. Furthermore, for machine learning algorithms to work correctly, the target classes must also be transformed into numerical representations using a Label Encoder.

The dataset split into 80% for training the models and the other 20% for testing. To address potential imbalances in class distribution, the training data underwent a process employing SMOTE (Synthetic Minority Over-Sampling Technique) This was done to ensure the class proportions in most cases.

To keep away from bias in the unbalanced dataset we implement Oversampling techniques that produced 500 samples for each of five classes. The balancing scheme represents better machine learning model performance by stopping model bias towards majority classes which leads to enhanced model generalization capabilities for all classification categories. Figure 2 outlines a classification framework that performs better in terms of generalization and exhibits more interpretable results.

DOI: http://dx.doi.org/10.25098/9.1.32

The Scientific Journal of Cihan University – Sulaimaniya **PP: 140-157**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

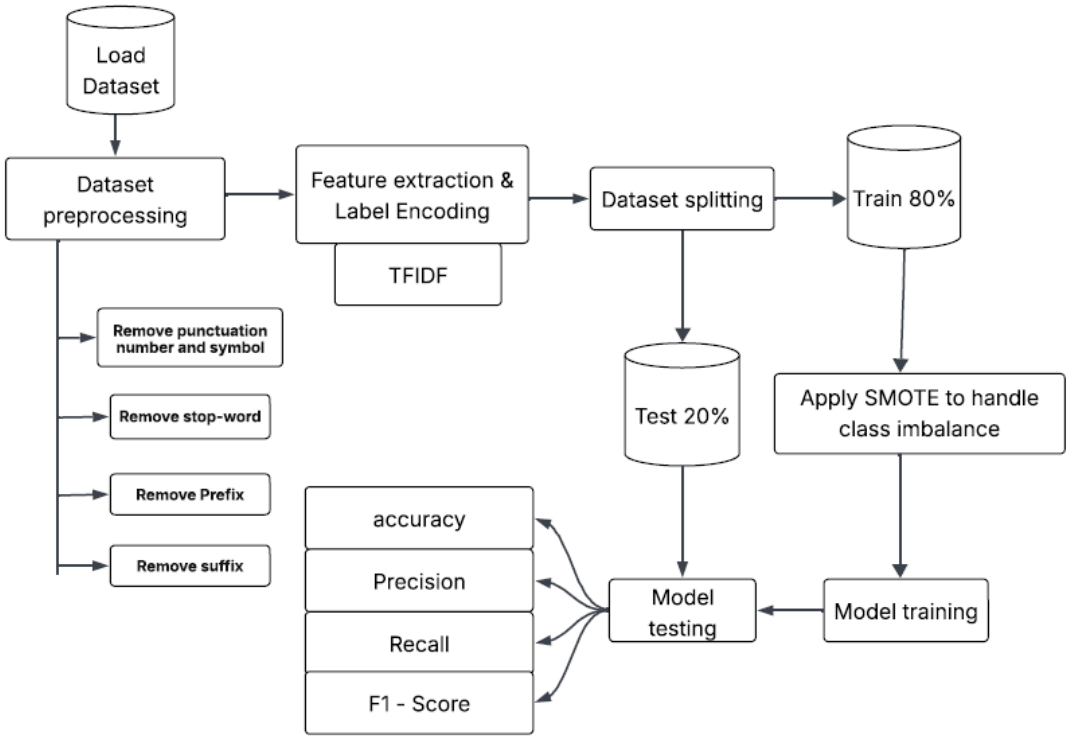

Figure 2. Methodology Block Diagram

4.1 Data collection

In this study 55883 headlines are analyzed. These headlines originated from several Kurdish web platforms – Xandan, Rudaw, NRT and Awena. The dataset contains an even distribution across five primary categories, specifically covering news related to technology, sports, health and business as shown in table 2.

Table 3. shows the number of different categories.

Table 2: Number of collected data according to web platforms

Website	Number of Samples
K24 www.xendan.org	12868
Rwdaw www.rudaw.net	8257
NRT www.nrttv.com	7541
politicpress www. politicpress.com	10767
Pukmedia www.pukmedia.com	3424
Xelk www.xelk.org	13026

DOI: http://dx.doi.org/10.25098/9.1.32

The Scientific Journal of Cihan University – Sulaimaniya **PP: 140-157**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Table 3: Number of collected data according to categories

Category	Number of samplers
Sports	15306
Health	11803
Business	13735
Social	9086
Technology	5953

4.2 Data Descriptions

Sorani Kurdish, much like Persian, Arabic, and Urdu, while the alphabets share a visible resemblance across these languages, differences in their Unicode employs and right-to-left writing system without capitalize form and grammatical gender. Sorani Kurdish alphabet has 36 letters divided into constant and vowels as shown in table 4.

Table 4: Kurdish Sorani vowels and consonant Letters.

Constant	ئ - ب - پ - ت - ج - چ - ح - خ - د - ر - ڕ - ز - ژ - س - ش - ع - غ - ف - ڤ - ق - ک - گ - ل - ڵ - م - ن - ه - و - ی
Constant in Latin	y - w - h - n - m - ll - l - k - g - q - v - f - x - ë - ş - s - j - z - rr - r - d - x - h – ç - c - t - p - b – a
Vowels	ێ، ی، وو، ۆ، و، ە، ا
Vowels in Latin	a-e -u-o- û-i- ë

Sorani Kurdish got both past and present tenses and singular and plural forms. Table 5 shows some sentences samples
for different Categories

Table 5: Samples of Kurdish Sorani Sentences.

Category	Text Sample	Translation
Sports	لێدوانی میسی و ڕامۆس بۆ هاندەران	Messi and Ramos speak to fans
Health	ڕۆژانە پێویستە چەند کوپ قاوە زیاتر نەخورێتەوە؟	How many cups of coffee is it advisable not to exceed per day
Business	بازرگانەکان کێبڕکێ دەکەن و بە کەمترین قازانج کاڵاکان دەفرۆشن	Business owners compete with another and market their products with the minimal possible profit
Social	پیاوێك چۆن لە ژێر سکەی شەمەندەفەرێکدا لە مردن گەڕایەوە؟	How did a man return from the brink of death after being underneath a train?
Technology	مەسنجەر گۆڕانکاری نوێ دەکات	Messenger is undergoing a new update

DOI: http://dx.doi.org/10.25098/9.1.32

The Scientific Journal of Cihan University – Sulaimaniya **PP: 140-157**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

4.3 Data preprocessing

many processes should be done for making data ready for machine learning:

- **Remove punctuation and symbol:** For reducing noise should be remove non-alphabetic characters.

In our dataset we removed (42) punctuation and symbol
[``"؛– X‹$“”(),".?<>:;|123456789-+*&^!%#@\/{}؟]

- **Normalization:** The text convert into a unified Kurdish script to get over inconsistencies in spelling and encoding.
- **Remove stop-word:** Kurdish stop-word like (لە ،بۆ ،کە…e.g.) must be remove to remain the meaningful content that contribute to classification accuracy.

in our dataset we defined (1224) stopwords.

- **Prefix removal:** are the word parts added to the beginning of a word to modify its meaning, Kurdish language like other language use prefixes to represent many grammatical functions, such as tense, negation, or modification the example for prefixes like (نـ, بە), the purpose for removing prefixes allowing the model to focus on the word's root meaning.
- **suffix removal:** is a group of letters added to the *end* of a word to change its meaning or grammatical function. These are very common in paperwork to indicate actions, states, roles, and qualities. is the process for removing prefixes (e.g., بە ,نـ) and suffixes (e.g.,ـەکان ,ـان) to return the word to the base or root form, this process known as stemming.
- **"Steaming"** is generally a standalone word with meanings related to steam or anger and isn't a typical prefix or suffix in paperwork terminology.

These steps improve the quality of data and aimed to standardize the input data for the machine learning models.

4.4 Machine Learning Models Training Parameters

Table 6 shows different hyper parameter tested for XGBoost, Random Forest, Support Vector Machine (SVM) together with LightGBM (LGBM) models. Choosing optimal values during model development isn't about just performance but also about finding the spot between predictive accuracy and operational efficiency.

DOI: http://dx.doi.org/10.25098/9.1.32

The Scientific Journal of Cihan University – Sulaimaniya **PP: 140-157**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Table 6. Machine Learning Model Training Parameters

Model	Parameters	Values
XGBoost	random_state	42
	n_estimators	100
	learning_rate	0.3
	max_depth	6
Random Forest	random_state	42
	n_estimators	100
	max_depth	10
SVM	random_state	42
	kernel	Linear
	probability	TRUE
LGBM	n_estimators	100
	learning_rate	0.1
	random_state	42

5. Results and Discussion

Four machine learning models (Random Forest, SVM, XGBoost and LGBM) multi-class classification implemented as shown in the presented confusion matrices in Figure (y1). The result of predicting five different classifications (sport, health, technology, business and social) are represented in each confusion matrix. The accurate classifications lined up along the matrix's diagonal. Meanwhile, any errors in classification will show up elsewhere in the matrix, scattered across the remaining cells. The darkness of its shades, essentially reflects event frequency. Darker colors indicate instances where a higher intensity corresponds to more frequent occurrences. The model performance in class discrimination is possible through matrix comparisons. This assists in choosing the best model to perform classification function. Looking at the confusion matrix data, it's clear the model's accuracy varies across different categories. SVM and XGBoost, are more strong classifiers, especially when it comes to business and social categories in comparison with LGBN and Random Forest.

DOI: http://dx.doi.org/10.25098/9.1.32

The Scientific Journal of Cihan University – Sulaimaniya **PP: 140-157**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

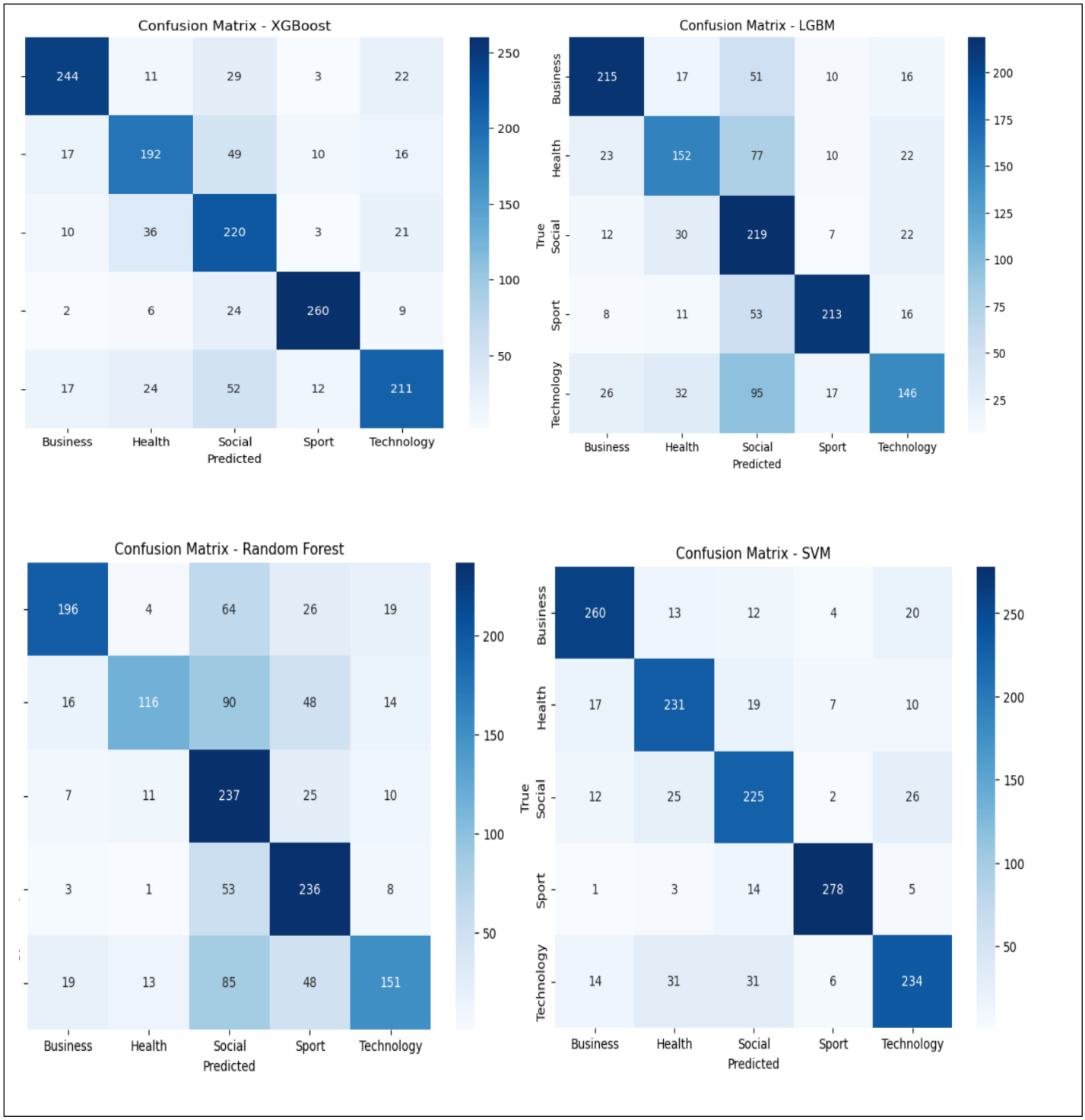

Figure 3. Machine Learning Models Confusion Matrix

When scrutinizing the performances of four distinct machine learning models Support Vector Machines (SVM), LightGBM (LGBM), XGBoost, and Random Forest it becomes clear that each models its own set of strengths and weaknesses under varying classification conditions. Assessing how well the models identified the given dataset was achieved using a base of performance metrics. This included accuracy and precision together with recall and F1-score. When comparing models,

DOI: http://dx.doi.org/10.25098/9.1.32

The Scientific Journal of Cihan University – Sulaimaniya **PP: 140-157**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

SVM often shows superior results. Its classification scores tend to be quite high, often surpassing XGBoost and LGBN, which generally perform well too. While Random Forest can still be useful, its performance is usually a bit lower than others. This can sometimes lead to less accurate classifications, particularly affecting how consistently it classifies different categories.

Table 7 represents different machine learning models metric results for different classes. SVM stands as the most effective model because it demonstrates peak accuracy across all groups including its exceptional F1-score of 0.93 in the culture class. model shows strong performance metrics in technology, commerce, social and sport categories because it maintains precision and recall levels above 0.75 and it is clear that all precision and recall values are close together SVM demonstrates exceptional reliability because it accurately predicts text classification. The XGBoost algorithm demonstrates consistent consistency throughout all categories reaching accuracy scores of 0.75. The culture class exhibits exceptional precision at 0.90 which guarantees highly accurate positive predictions for this category. The strong results obtained by XGBoost demonstrate its effectiveness when used to identify differences between technology and commerce classes.

The robust system shows strength in general but needs additional adjustments to enhance its ability to identify social and maintain consistent performance in those particular areas. The Random Forest algorithm displays performance ranges that are higher in variability between recall and precision. The class commerce precision 0.80 and its recall is 0.41 and the same gap appears for social and sport classes. The technology and culture F-1 score remain shows to be higher than the other classes. The model demonstrates poor performance in comparison with the other models. The LBGN algorithm displays performance ranges that are higher in variability between recall and precision in time these variabilities are less than random forest. The class culture precision 0.83 and its recall is 0.71. there is a large variability in sport class and the same difference between precision and recall presented in social class in comparison with the other classes.

DOI: http://dx.doi.org/10.25098/9.1.32

The Scientific Journal of Cihan University – Sulaimaniya **PP: 140-157**
Volume (9), Issue (1), June 2025
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Table 7. Machine Learning Models Results

Models	Class	accuracy	precision	recall	F1-score
Random forest	Technology	0.62	0.80	0.64	0.71
	Commerce		0.80	0.41	0.54
	Social		0.44	0.81	0.57
	Culture		0.62	0.78	0.69
	Sport		0.74	0.48	0.58
SVM	Technology	0.82	0.86	0.84	0.85
	Commerce		0.76	0.81	0.79
	Social		0.75	0.78	0.76
	Culture		0.94	0.92	0.93
	Sport		0.79	0.74	0.77
LGBN	Technology	0.63	0.76	0.70	0.73
	Commerce		0.63	0.54	0.58
	Social		0.44	0.76	0.56
	Culture		0.83	0.71	0.76
	Sport		0.66	0.46	0.54
XGBoost	Technology	0.75	0.84	0.79	0.81
	Commerce		0.71	0.68	0.69
	Social		0.59	0.76	0.66
	Culture		0.90	0.86	0.88
	Sport		0.76	0.67	0.71

6. Conclusions

The research explores numerous machines learning for text classification identification by collecting data from different Kurdish news sites. SVM shows the most effective performance since it achieves the highest accuracy scores and F1-scores across all classification categories. The XGBoost algorithms deliver good performance in culture and technology categorization yet Random Forest and LGBN fails to achieve accurate results for social class. Modern machine learning models have proven their ability to boost text classification operations according to research evidence. The research progression will focus on achieving better model performance through tuning of hyper parameters and expansion of datasets along with additional feature implementation to address misclassification problems and boost precision.

DOI: http://dx.doi.org/10.25098/9.1.32

References

[1] K. M. Awlla, H. Veisi & A. A. Abdullah. " Sentiment analysis in low-resource contexts: BERT's impact on Central Kurdish. Language Resources and Evaluation ". springer, available at: https://www.scilit.com/publications/3c7581659f5c9f6424340b1709f55970. 2025.

[2] N. A. Atadoga, E. O. Sodiya, U. J. Umoga, & O. O. Amoo. "A comprehensive review of machine learning's role in enhancing network security and threat detection". World Journal of Advanced Research and Reviews, vol.21, no.2. pp877-886, 2024.

[3] K. Taha, P. D. Yoo, C. Yeun, D. Homouz & A. Taha. "A comprehensive survey of text classification techniques and their research applications: Observational and experimental insights". Computer Science Review, vol.54, 2024.

[4] A. M Saeed, S. Badawi, S.A. Ahmed, & D. A Hassan. "Comparison of feature selection methods in Kurdish text classification". Iran Journal of Computer Science. Vol.7, no.1, pp 55-64, 2023.

[5] H. Allam, L. Makubvure, B. Gyamfi, K. N. Graham, & K. Akinwolere. "Text classification: How machine learning is revolutionizing text categorization". Vol.16, no.2, 2025.

[6] A. A. Abdullah, S. S. Muhamad, & H. Veisi. "Enhancing Kurdish Text-to-Speech with Native Corpus Training: A High-Quality WaveGlow Vocoder Approach", 2024.

[7] S. S. Badawi. Bridging the gap. ARO-The Scientific Journal of Koya University. Vol.12, no.1,pp.100-107, 2024.

[8] H. A. Ahmad,T. A. Rashid. Central Kurdish Text-to-Speech Synthesis with Novel End-to-End Transformer Training. Algorithms, vol.17, no.7, pp.2-19, 2024.

[9] K. M. H. Rawf, S. H. T. Karim, A. O., Abdulrahman& K. J Ghafoor. Dataset for the recognition of Kurdish sound dialects. Data in Brief, Vol.53, 2024.

[10] A. M. Saeed. "AN AUTOMATED NEW APPROACH IN FAST TEXT CLASSIFICATION: a CASE STUDY FOR KURDISH TEXT". Science Journal of University of Zakho, vol.12, no.3, pp. 329-335, 2024.

[11] SH. M. Shareef, & A.M. Ali. "Deep learning-based digitization of Kurdish text handwritten in the e-government system". Indonesian Journal of Electrical Engineering and Computer Science, vol.35, no.3, pp.1865-1875, 2024.

[12] H.A. Ahmad, & T.A. Rashid. "Gigant-KTTS dataset: Towards building an extensive gigant dataset for Kurdish text-to-speech systems". Data in Brief,vol 55, 2024.

[13] A. A. Abdullah, S. H. Abdulla, D.M. Toufiq, H. S. Maghdid, T. A. Rashid, P. F. Farho, S.S Sabr, A. H. Taher, D. S. Hamad, H. Veisi & A.T. Asaad. " NER- RoBERTa: Fine-Tuning RoBERTa for Named Entity Recognition (NER) within low-resource languages". arXiv, 2024.

[14] D. Mahmud, B. A. Abdalla & A. Faraj. "Twitter Sentiment analysis for Kurdish language. Qalaai Zanist Scientific Journal". vol. 8, no.4, pp. 1132-1144, 2023.

[15] S. Badawi. "KurdSum: A new benchmark dataset for the Kurdish text summarization". Natural Language Processing Journal, vol. 5, 2023.

[16] S. Badawi, A. M. Saeed, S. A. Ahmed, P. A. Abdalla & D.A. Hassan. " Kurdish News Dataset Headlines (KNDH) through multiclass classification". Data in Brief,vol. 48, 2023.

DOI: http://dx.doi.org/10.25098/9.1.32

[17] R. Filippis, A. Al Foysal. "Predicting Bipolar Disorder Treatment Outcomes with Machine Learning: A Comprehensive Evaluation of Random Forest, Gradient Boosting, and Ensemble Approaches". . Open Access Library Journal, vol.12, no.2, 1–18, 2025.

[18] D. L. Garcia, B. J. Kotzian, J. Yang, B. Mwangi, B. Cao, L. N. P. Lima, M. B. Bermudez, M. V. Boeira, F. Kapczinski, I. C. Passos. "The impact of machine learning techniques in the study of
bipolar disorder: A systematic review". Neuroscience & Biobehavioral Reviews, vol. 80, pp. 538-554, 2023.

[19] G. Khyathi, K. P. Indumathi, H. A. Jumana, F. J. M. Lisa, S. Siluvari, G. Krishnaprakash , Support
Vector Machines A Literature Review on Their Application in Analyzing Mass Data for Public Health,
Cureus, vol.17, no.1,2025.

[20] X. Zhang, Y. Wang, Z. Zhuang, Y. Liu, Ch Yuan, L. Su, J. Shaou & P. W. Chan. "Comparison of simulating visibility using XGBoost and IMPROVE method: a case study in East China". Forntiers in Environmental Science, vol.12, 2025.

[21] Kh. A. Ben Hamou, Zahi Jarir, Selwa Elfirdoussi. "Application of LightGBM Algorithm in Production Scheduling Optimization on Non-Identical Parallel Machines". Engineering, Technology & Applied Science Research. Vol. 14,no. 6 , pp. 17973-17978, 2024.

[22] N. Klingler. "Confusion Matrix in Machine Learning – A complete guide (2025)". Available at : https://viso.ai/deep-learning/confusion-matrix/, 2024

DOI: http://dx.doi.org/10.25098/9.1.32