


Research Article


Arabic Crime Tweet Filtering and Prediction Using Machine Learning

Zainab Khyioon Abdalrdha^{1,*} 

Informatics Institute of Postgraduate
Studies, Iraqi Commission for Computers &
Informatics
Baghdad, Iraq
phd202120695@iips.edu.iq

Prof. Dr. Abbas Mohsin Al-Bakry² 

University of Information
Technology and Communication
Baghdad, Iraq
abbasm.albakry@uoitc.edu.iq

Prof. Dr. Alaa K. Farhan³ 

University of Technology
Department of Computer Sciences
Baghdad, Iraq
110030@uotechnology.edu.iq

ARTICLE INFO

Article History

Received: 01/03/2023

Accepted: 11/04/2024

Published: 01/06/2024

This is an open-access article under the
CC BY 4.0 license:

<http://creativecommons.org/licenses/by/4.0/>



ABSTRACT

Crime is undeniably rising, thus negatively affecting countries' economies. Despite several efforts to study crime prediction to reduce crime rates, few studies take the timeline factor into account when extracting crime-related tweets to predict crime. Aiming to predict Arabic crime tweets on Twitter/X, this study predicts crimes after analyzing social sentiment—that is, whether a tweet raises positive, negative, or neutral feelings—and filters the tweets based on crime behavior through an intelligent dictionary built through a genetic algorithm. The study uses a variety of machine learning (ML) models—random forest, logistic regression, and decision trees—which are assessed according to their accuracy, precision, recall, and F1 scores to guarantee robustness and dependability in crime prediction. The accuracy after filtering crimes based on an intelligent dictionary are 97% for decision tree, 97% for random forest, and 94.43% for logistic regression. This research will provide insight into potential crime attitudes and

Keywords: *Cybercrime, Machine Learning, Twitter Analysis, Natural Language Processing (NLP), Random Forest, Logistic Regression*

1. INTRODUCTION

Social media serves various purposes but can also facilitate crimes [1]. Sharing personal information online can result in criminal activities. Victims might hesitate to report crimes because they consider them insignificant, feel embarrassed, or are unaware of the process. Social media monitoring can enhance traditional crime reporting. Social media is utilized to enable illegal activities, similar to other new technology and communication platforms [2]. Protecting private data during network transmissions is crucial [3]. Twitter is distinguished from other social networking sites by the fact that it allows users to submit news, thoughts, and ideas under a 280-character limit. In contrast to tweets, text relationships do not have the same method for sharing information [4][5]. Data dumps, security lapses, ransomware, vulnerabilities, DDoS assaults, zero-day exploits, and public events are a few instances of cyber threats that are regularly spoken about on Twitter, a platform heavily utilized for this kind of activity [6]. Researchers can use Twitter's tracking and tweeting capabilities to assess interest in specific topics and uncover unforeseen cyber threats in real time [7]. Security intelligence utilizes artificial intelligence to gather and structure information related to cyber dangers [8]. Conventional machine learning (ML) algorithms have demonstrated efficacy in predicting crime. Several methods, including decision tree (DT), logistic regression (LR), and random forest (RF), have been used to analyze crime data to detect trends for predicting criminal behavior. Traditional ML models require less data and are simpler to interpret compared with deep learning, which depends on extensive data and intricate neural networks. This research suggests a method for forecasting Arabic crime tweets by utilizing ML models to predict keywords in criminal Twitter datasets from 18,493 collected tweets [9]. This work predicted Arabic crime tweets by using ML models based on various keywords that contain seven classes ('إرهاب', 'terrorism', 'تهريب المخدرات', 'drug smuggling', 'تنظيم داعش', 'ISIS', 'تتمر', 'bullying', 'تهريب السلاح', 'arms smuggling', 'سرقة بنك', 'bank robbery', 'قتل', 'murder'). These seven classes are used for predicting crimes after filtering datasets using genetic algorithms, and the prediction performance



is evaluated. The remainder of this paper is organized as follows: Section 2 covers relevant related literature. Section 3 outlines the methodology. Section 4 details the experiments and their outcomes. Section 5 concludes the evaluation and proposes potential areas for future investigation.

2. RELATED WORK

Crime is prevalent globally. Hence, law enforcement agencies are seeking sophisticated information systems to reduce crime rates and safeguard society. Criminology is the systematic examination of criminal behavior to identify the underlying reasons for crimes through the collection and analysis of data. Research on forecasting criminal activity in Arabic tweets is scarce. This section delineates some of the existing studies. Sentiment analysis was applied by the authors of [10] to monitor instances of criminal behavior on Twitter. Brown clustering, which was applied to the analysis of a large array of unlabeled tweets, produced superior results in anticipating crime rates when compared with traditional data collection approaches. From 2014 to 2018, a study conducted on Twitter examined crime rates in seven cities in India, and the results revealed a detection accuracy of 70%. Ghaziabad had the greatest crime rate, whereas Jammu had the lowest crime rate during that period. The study emphasizes the need to make a distinction between cities with high crime and cities with low crime.

The authors of [11] performed text mining-based categorization by using naïve Bayesian, RF, J48, and ZeroR. Pre-processing data can help improve categorization accuracy. The Twitter real-time dataset that they used contained 500 tweets, with 270 offenders and 230 non-criminals. Four classifiers were tested using a sizable dataset for accuracy, precision, recall, and F1 score. With an accuracy percentage of 98.1%, RF was the most accurate, while ZeroR was the least accurate at 61.5%. Unbalanced datasets were addressed using receiver operating characteristic curve (ROC) analysis, highlighting the necessity of choosing the right classifier for precise predictions. Sh. M. M. Matias and collaborators [12] created an ML method to forecast cybercrime. They did not evaluate parts of speech tags. By concentrating on preprocessing and sentiment analysis and applying techniques such as naïve Bayes, DTs, and support vector machines (SVMs), researchers examined cyberbullying and cyber dangers. A total of 25% of the data were used for testing after the model had been trained on 75% of the data. Using Twitter data to predict large-scale social media crimes is advisable according to Sr. C M [13].

Social media data analysis is used to predict feelings and identify offenders. The approach forecasts cyberbullying, cyberstalking, online fraud, cyber harassment, and cyberbanking using social media data. This research uses multinomial naïve Bayes (MNB), KNN, or SVM to categorize tweets. The three algorithms had precision, recall, and F-measure scores of more than 0.9, according to the data. Using a combination of lexicon-based and deep learning techniques, the hybrid method presented by M. B. M. Azizi [14] employed BERT as the DL model to categorize and detect crimes. This model included softmax prediction, padding, attention masks, and special tokens and is nearly as good at solving complicated problems as the brain. The research looked at more than 70,000 tweets, comprising 43,000 regular tweets and 27,000 tweets related to crimes. With the use of Twitter's streaming API, the tweets were divided into 18,000 tweets on crimes and 37,000 tweets about general topics. The F1-score, classification accuracy, recall accuracy, loss, precision, and precision score of the suggested Twitter crime detection technique were 94.92, 94.91%, 16.26%, and 94.91%, respectively. The suggested method might be used to detect cybercrime on social media. Aditi, Parth, et al. [15] created a Twitter-based algorithm that uses tweets about attacks to identify terrorist acts. To determine tweet types, they employed a ternary search method in conjunction with Aho-Corasick, using 1,000 and 250 tweets' worth of Twitter 4j API data in the investigation. The actual data showed a higher percentage of catastrophic terrorist incidents than anticipated by the KNN and SVM algorithms. ML is used in the suggested method to locate places affected by COVID-19.

The authors in [16] utilized the AraNews dataset to train the models. The study employed the term frequency-inverse document frequency (TF-IDF) approach for feature extraction. Subsequently, three ML methods were employed to forecast the occurrence of fake news: RF classifier, naïve Bayes, and LR. The RF model attained the highest accuracy. The authors [17] offered a method for enhancing Arabic fake news prediction algorithms. To increase prediction accuracy, this research uses text, user, and content aspects to combat disinformation and the quick spread of fake news on social media platforms. The technique converts Twitter content into features and selects high-ranking features by using TF-IDF. Relevant user attributes are found using a fuzzy model. RF is updated to outperform other ML techniques. The paper demonstrates how Arabic fake news can be identified with 0.895 accuracy using the improved RF model. In contrast, the naïve Bayesian and SVM methods achieved accuracies of 0.809 and 0.848, respectively. Thus, Arabic fake news may be reliably predicted by the method that utilizes modified RF and fuzzy logic.

3. THE PROPOSED METHOD FOR PREDICTING CRIMES

Forecasting Arabic criminal activity on Twitter using ML algorithms requires multiple key stages, as seen in Fig. 1.

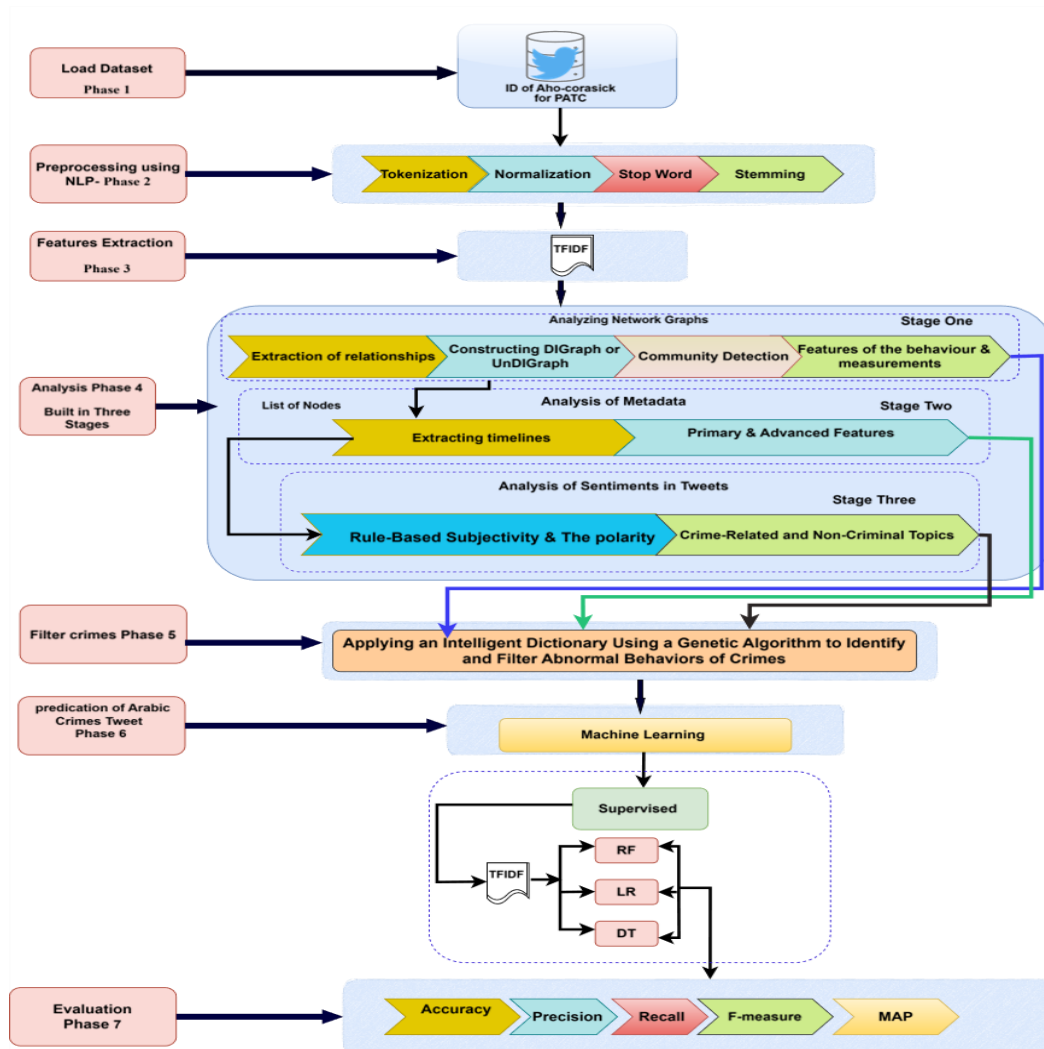


Fig. 1. The proposed method for predicting crimes in Arabic tweets.

The stages described in the methodology of the proposed method are explained below in detail.

A. Arabic crime Twitter dataset:

Although the Twitter development team offers an official Twitter API [18], Python has several library tools for data extraction, including Tweepy and Scrape. This study uses a dataset constructed using an intelligent lexicon [9] to extract and investigate criminal activity, analyze tweets, and filter out anomalous behavior by using an intelligent dictionary constructed using a genetic algorithm [19].

B. Dataset preprocessing

The research employs preprocessing techniques such as normalization, tokenization, stop-word removal, and stemming to prepare a dataset of 18,493 tweets for crime tweet prediction. Consistent categorization is ensured across all datasets by performing normalization to lower noise and applying tokenization to turn characters into tokens for linguistic analysis.

This study utilized NLTK, a renowned Python library for natural language processing that has a wide range of functions for encoding words and sentences, such as `nlk.word_tokenize()` and `nlk.sent_tokenize()` for word and sentence encoding, respectively. Stop-word removal eliminates superfluous words, and stemming eliminates prefixes and suffixes from inflected words [20]. Table I shows examples of stop-words.

TABLE I. SOME EXAMPLES OF STOP-WORDS.

| No. Stop words | Stop words Arabic | No. Stop words | Stop words Arabic | No. Stop words | Stop words Arabic |
|----------------|-------------------|----------------|-------------------|----------------|-------------------|
| 1 | 'إذ' | 14 | 'مهما' | 27 | 'هذه' |
| 2 | 'لسيما' | 15 | 'ليت' | 28 | 'اللتان' |
| 3 | 'أنتم' | 16 | 'لسنا' | 29 | 'لسن' |
| 4 | 'لدى' | 17 | 'إلا' | 30 | 'له' |
| 5 | 'إذما' | 18 | 'لعل' | 31 | 'إليك' |
| 6 | 'لست' | 19 | 'التي' | 32 | 'لها' |
| 7 | 'إذن' | 20 | 'لك' | 33 | 'إليكم' |
| 8 | 'لستم' | 21 | 'الذي' | 34 | 'اللواتي' |
| 9 | 'أف' | 22 | 'لكم' | 35 | 'لنا' |
| 10 | 'لستما' | 23 | 'الذين' | 36 | 'إلى' |
| 11 | 'أقل' | 24 | 'لكما' | 37 | 'أينما' |
| 12 | 'لستن' | 25 | 'اللاتي' | 38 | 'اللاتي' |
| 13 | 'أكثر' | 26 | 'لكن' | 39 | 'لكنما' |

Arabic texts have many stop words that add significantly to their meaning. These frequently occurring words, which are inherently common, can be eliminated during the pre-processing of texts due to their repetitiveness and lack of significant meaning in the phrase. Hence, in certain instances, the presence of omitted words, particularly when they are eliminated, can significantly impact the sentence's meaning and the comprehension of its context, as they play a crucial role in communication. The sentence's significance resides primarily in responses or tweets, with negative terms like "["الن", "الم", "ما", "لا"]" serving as an example.

C. Phases of analysis

In this section, the critical steps for analyzing and filtering crimes are completely completed and published according to the paper [19], which included several stages consisting of three parts: the network analysis stage, the metadata stage, and the sentiment analysis stage.

D. Identifying and filtering abnormal Arabic crime using an intelligent dictionary

The next step involves implementing a filtering method to identify and detect abnormal behavior in criminal activities. A genetic algorithm is utilized in this stage to create a dictionary of users with actions that are deemed highly intriguing. The search region is restricted by considering all data collected from the graph network and metadata analysis. The dataset comprises 18,493 tweets collected by the Aho-Corasick algorithm. This technique efficiently establishes a vocabulary for searching through a vast text corpus. A total of 15,485 tweets were chosen for examination in this study. The search area was then trimmed down to 3,228 profiles, indicating a 17.45% decrease.

A Twitter profile linked to illicit activities presents a significant danger. An intelligent dictionary was developed to eliminate unengaging profiles to decrease the input size of components in this investigation. Intelligent dictionaries are developed through the analysis of interactive graph network data and user behavioral patterns obtained from metadata. The key categories utilized for constructing the intelligent dictionary using the evolutionary algorithm are "old spreader," "influencer (I)," "spreader (RT)," "influencer (II)," "constant spreader," "new profiles with significant engagement," and "influencers (III)." The component for detecting and filtering aberrant behavior finds individuals displaying suspicious conduct and refers them to the next component for further examination. This step entails scrutinizing their material and delving deeper into their posts. This step is executed in detail in [19], which examined crime prediction based on ML algorithms. Through an analysis shown in Fig. 2, the X-axis shows Twitter usernames, and the Y-axis shows user categories. The different behaviors, such as old spreaders, influencers (III), constant spreaders, and others, provide context for these categories. The Y-axis position represents the user's behavioral category, while the x-axis position indicates a Twitter user. The categories include "old spreaders," "influencers," "new profiles with high activity," "constant spreaders," and "others." Security agencies can use the method to identify hazardous disinformation spreaders on Twitter or reveal user activity patterns. A genetic algorithm, which is an optimization technique that draws inspiration from natural selection,

classifies these behaviors. Fig. 2 depicts the genetic algorithm-determined Twitter user behavior groupings. The abundance of data points and a broad variety of categories indicate a thorough user behavior analysis.

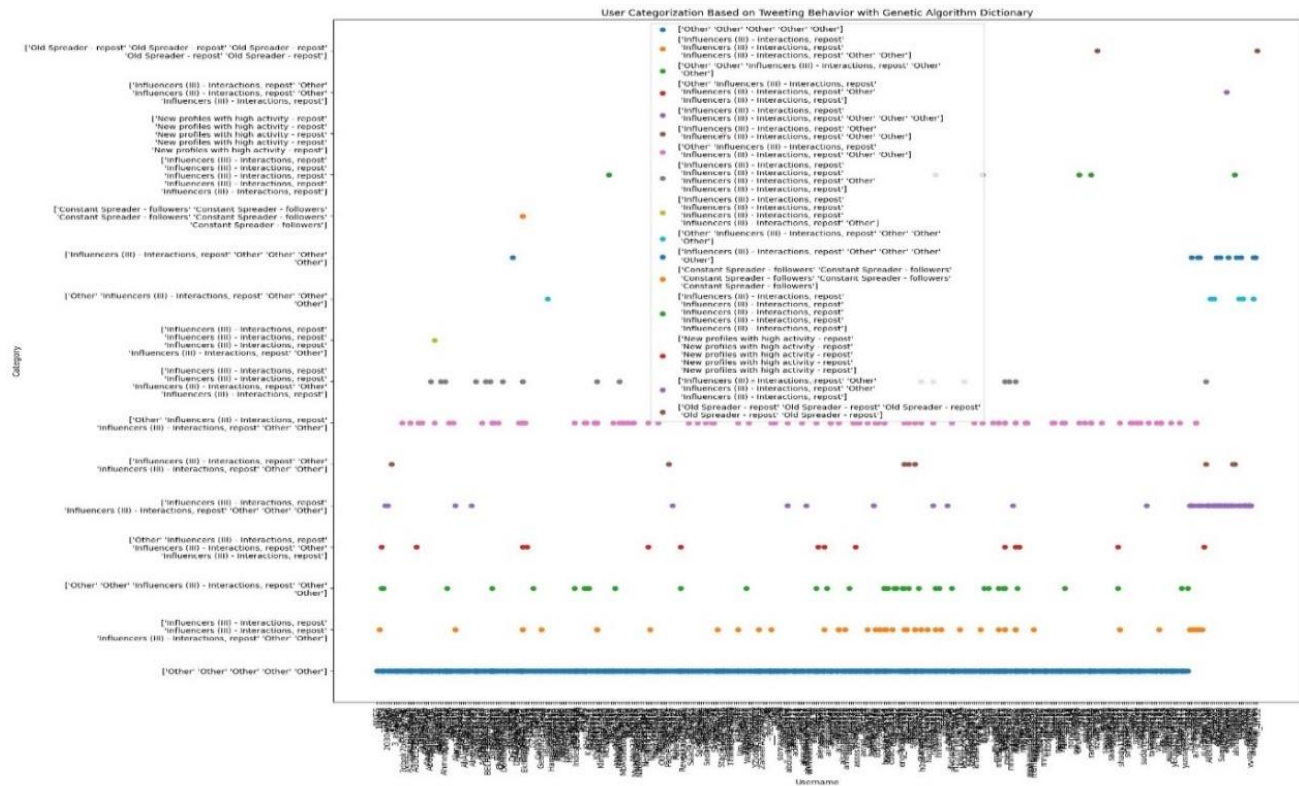


Fig. 2. Using a genetic algorithm dictionary to filter abnormal Arabic criminal behavior [19].

E. Prediction of Arabic crime tweets using machine learning

ML was used to forecast Arabic crime tweets in this section by using an algorithm. This study uses ML due to its ability to easily define parameters and achieve very accurate results. Thus, a comparison among ML approaches can be conducted. We employed three ML algorithms, namely, RF, LR, and DT. This study utilizes the Arabic crime Twitter dataset to investigate the impact of these algorithms on text classification. To predict Arabic crime tweets, this part focuses on feature extraction through the utilization of TF-IDF and the application of ML techniques.

- **Feature extraction**

Feature extraction aims to identify certain data elements in natural language documents that are used to train the classifiers. The documents typically consist of string characters. However, ML algorithms are unable to process such strings directly. Thus, these strings are transformed into a suitable format for ML classifiers [21]. This paper will discuss the TF-IDF approach.

- **TF-IDF feature extraction**

According to the TF weighting approach, weight is a numerical statistic that indicates how important a word is to document in a particular kind (class) within a corpus or collection. When a word appears frequently in one paper, it may carry greater significance. Excessive weight values may provide additional information for text prediction and classification. Equation (1) illustrates that TF-based high-weight values, however, may not always provide useful information for classification because words that are frequently found in a given corpus may not be highly significant [22]. The term's local frequency is divided by its document frequency using TF-IDF. By giving uncommon words more weight, this metric will draw attention to terms that are distinctive to particular authors. The provided word's IDF is computed using Equation (2) [23–24].

$$\text{TF}(t, d) = \text{count of } t \text{ in } d / \text{number of words in } d \quad (1)$$

$$\text{IDF}(t) = \text{occurrence of } t \text{ in documents}$$

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \log(N / (\text{DF}(t) + 1)) \quad (2)$$

IDF is a statistical measure used to determine the importance of a word within a dataset. The IDF method is integrated to reduce the weight of frequently occurring words in the text group and increase the weight of less common words [25].

1. Random forest algorithm

RF is a classification and regression technique that uses a group of people for learning. It involves building an array of DTs. RF uses two kinds of randomization.

a) The training data is divided into discrete subsets, called bootstrap samples, from which each DT is constructed.

b) At each node split in a DT, a random subset of m variables is selected from the original set of variables, and the optimal split is determined based on these m variables.

The forecasts of the trees that RF builds are combined for an unknown circumstance. In regression tasks, RF averages the results; for classification problems, it employs majority voting to determine the final prediction; and in regression tasks, it performs averaging [26].

2. Logistic regression classifier

The subsidiary variable in this classifier is binary, indicating that it only holds data. LR uses a function called sigmoid to model the relationship between an event and its dependence on several variables. This algorithm is a soft classification technique that provides a probability as the outcome rather than a binary classification. This study utilizes multinomial LR with the multiclass parameter setting in the LR class of scikit-learn, instead of the sigmoid function. This model uses the L-Broyden-Fletcher-Goldfarb-Shanno (LBFGG) algorithm, which is a limited-memory optimization algorithm that is suitable for large-parameter problems and uses gradients to find local maxima and minima. LBFGGs work efficiently for LR, supporting multi-class classification, L2 regularization, and stability in small to medium-sized datasets [26].

3. Decision tree

In ML, DTs are employed for classification and regression. This supervised learning approach iteratively uses the values of the independent variables to partition the data into smaller and smaller subgroups, much like a tree splitting into subtrees [27][28]. The DT chooses the feature that reduces data impurity the best at each split based on measures including information gain, Gini impurity, and entropy. Maximum tree depth, minimum number of samples in a leaf node, or minimum impurity reduction are all examples of stopping criteria that can be reached to end the process. A tree-like structure depicting a chain of decisions and their consequences is the end product. From the root node, the DT algorithm travels to the leaf node, where it uses the mean or majority class to make a forecast for a new data point. Referring to Fig. 2, whether the link between the two variables is linear or nonlinear, DTs can handle it with ease and clarity. Overfitting is a common problem with these models, but trimming, bagging, and RFs can address this issue [27]. Entropy and Gini metrics [27] are defined as

$$E(D) = -\sum_{i=1}^k P(C_i) * \log(P(C_i)) \quad (3)$$

where C_i is the classes, and P is the probability of the class [26]. Information gain (D , features) for a feature in dataset D is calculated as the difference between the entropy of the original dataset and the weighted sum of entropies of subsets created by splitting on that feature, as illustrated in Equation (4)

$$IG(D, features) = E(D) - \sum \frac{(v \text{ in features})|D_v|}{|D|} * E(D_v) \quad (4)$$

where v is the iteration over the possible values of the feature, D_v is the subset of data points in D having the feature value v , and $|D|$ represents the total number of data points in D .

The method selects features that maximize information gain, thus minimizing entropy at each node. This recursive process continues until certain stopping conditions are met, such as reaching a predefined maximum depth or having a minimum number of samples in a leaf node [29].

4. RESULT AND DISCUSSION

ML algorithms show promise in predicting criminal behavior by utilizing extensive datasets and complex algorithms to enhance accuracy and efficacy. The algorithms can examine extensive datasets and detect trends in criminal activity by analyzing substantial volumes of data from social media networks. The results of predicting Arabic crime tweets using DTs, LR, and RFs are shown in Table 2. After a prediction model is built using training data, the accuracy of the learned model prediction must be calculated using test data to see how it will perform on future data. The confusion matrix is used in several classifier accuracy metrics in the literature. The confusion matrix is a table with dimensions of m by m , where m is the number of classes, and is used to evaluate the trained classifier's capacity to distinguish samples of multiple classes using actual and predicted class labels. Equations 5, 6, 7, and 8 provide the performance measurements [30].

$$\text{Precision} = \frac{TP}{TP + FP} \quad (5)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (6)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

$$\text{F1score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

Another performance metric is the Matthews correlation coefficient (MCC), a contingency matrix method for determining the Pearson product-moment correlation coefficient between actual and predicted values. It is an alternate metric unaffected by the imbalanced dataset issue. With regard to M 's entries, MCC is given as

$$\text{MCC} = \frac{TP \cdot TN - FN \cdot FP}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \quad (9)$$

(Optimal value: +1); lowest value: -1

When the binary predictor accurately predicts the majority of positive data instances and the majority of negative data instances, MCC is the only binary classification rate that yields a high score [31]. In evaluating models that are used for classification or prediction, the area under the ROC curve (AUC)-ROC performance metric can be used because it can distinguish between positive and negative events at different levels of class for classification. In addition, it is useful for operating the ROC because the AUC-ROC metric can assess the quality of the economic values of the classifier. This metric is represented by comparing TPR with FPR. In the case of evaluating the model, confusion matrices show both TPR and FPR. TPR is defined as the ratio of true positive cases correctly anticipated by a test to the total number of true positive cases, while FPR is the number of true negatives misclassified as positive by the test to the total number of true negative cases. Ideally, the values for AUC-ROC must be more than 0 and TPR hold 100% while FPR holds 0%.

The efficiency of the proposed methodology is higher than the efficiency of random classification. Thus, a greater AUC signifies the ability of a model to classify successfully between the positive and negative categories. An AUC value of 0.5 indicates that the classifier is performing randomly. An AUC of less than 0.5 indicates an inverted classifier. If the predictions of a model conflict with the actual labels, then the model's performance is inferior to that of random classification. The AUC-ROC is significant because it provides a comprehensive summary of model performance across different categorization thresholds. This metric is especially beneficial for datasets that have an unequal distribution of classes or where the costs of false positive and false negative predictions vary [32]. Another performance metrics is mean average precision (mAP). Data on precision and recall are gathered for each class throughout the preceding step. The objective of the present phase is to compute the average precision (AP) for each class

and then calculate the mean to achieve the mAP. The AP can be computed either by utilizing all the points on the precision–recall curve or by selecting selected points that correspond to a series of evenly spaced recall levels ranging from 0 to 1 [33].

TABLE II. PERFORMANCE RESULTS FOR PREDICTING CRIMES IN ARABIC TWEETS.

| Model for the Proposed Method | Precision | Recall | F1-score | Accuracy | Support |
|---|-----------|--------|----------|----------|---------|
| Before sorting dataset data time RF + TF- IDF | 94.94% | 94.62% | 94.65% | 95% | 521 |
| After sorting dataset data time RF + TF- IDF | 96.81% | 96.73% | 96.75% | 97% | 521 |
| Before sorting dataset data time LR +TF- IDF | 94.13% | 93.47% | 92.93% | 93.47% | 521 |
| After sorting dataset data time LR +TF- IDF | 94.74% | 94.43% | 94.13% | 94.43% | 521 |
| Before sorting dataset data time DT +TF- IDF | 96.14% | 95.96% | 95.99% | 96% | 521 |
| After sorting dataset data time DT +TF- IDF | 96.77% | 96.73% | 96.75% | 97% | 521 |

The result shown in Table II compares the performance of the three ML models using a proposed method on an imbalanced dataset that changes after applying a filtering method to identify and detect abnormal behavior in criminal activities using the genetic algorithm from the balanced dataset to the imbalanced. This paper uses an Arabic tweet crime prediction that was subjected to TF-IDF vectorization before and after sorting timeline extraction. An analysis of the results for RF, LR, and DT indicate an enhancement in performance after sorting the timeline. Organizing tweets chronologically can enhance prediction algorithms by maintaining the temporal context and sequence of events, facilitating a more accurate analysis of trends and changes in conversations. It enables the development of time-dependent characteristics that improve model forecasts. Studying tweets in chronological order facilitates cohort analysis, which helps in monitoring the behavior or mood of particular groups over time. It helps identify patterns, spikes, or decreases in activity, allowing predictive models to anticipate future trends. Organizing tweets also enables more precise causal inference and more accurate evaluations of model performance when dividing data into training and testing sets. When these models are applied, RF shows the highest improvement after sorting the timeline, with its accuracy increasing from 95% to 97%. Similar increases in precision, recall, and F1-score were observed. LR shows minimal improvement after sorting, maintaining accuracy and recall, but its precision and F1-score increased slightly. Moreover, it is less influenced by temporal data sequences. DTs can use the temporal structure of the data to make more meaningful categorization, which may explain why their accuracy increased from 96% to 97% across all metrics. After sorting, the models worked better, the metric variance dropped, and the risk of overfitting due to noise in randomized order data decreased. To guarantee correct comparison of metrics such as recall and precision, the “Support” column, which indicates label occurrences, stays at 521 throughout all models and circumstances. The result is shown in Figures 3,4,5.

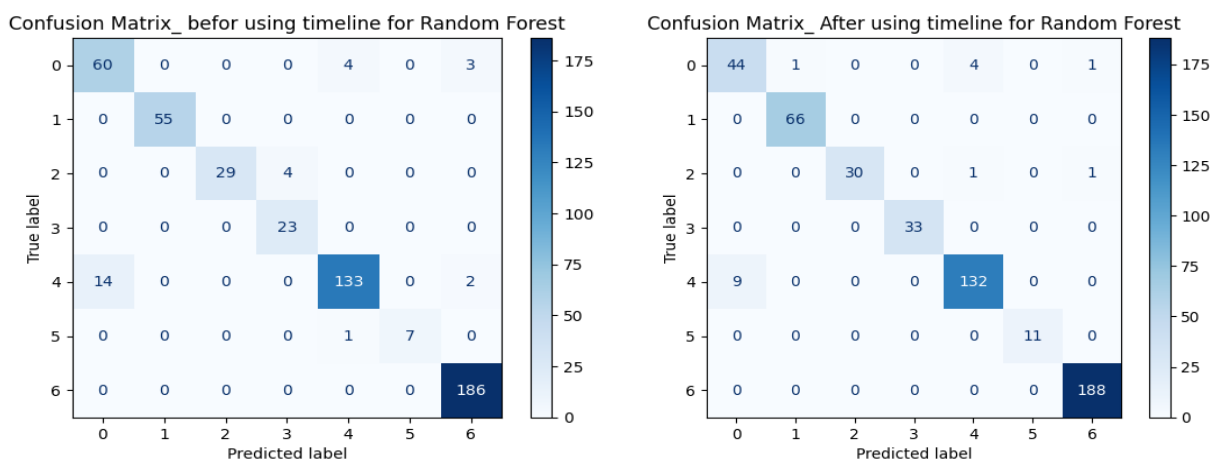


Fig. 3. Confusion matrix for Arabic tweet crime prediction for random forest.

The result in Fig. 3 for RF compares the model’s performance before and after sorting the dataset by timeline. The observed changes in the confusion matrix were analyzed. The sorting method of the dataset decreases off-diagonal

items in the confusion matrix, thereby reducing misclassifications and enhancing the model’s accuracy by considering the chronological order of tweets. Data sorting or an earlier representation of this class in the data could explain the rise in false negatives and the decrease in true positives for label 0 from 60 to 44. The model is more conservative in predicting when using a timeline. Label 4 experienced a reduction in false positives from 14 to 9, demonstrating an enhancement in its ability to differentiate from other classes. Label 5 misclassification decreased from 7 to 0, while false negatives rose from 1 to 11. Label 6 forecasts showed minimal enhancement, with the number of accurate positive results rising from 186 to 188. The presence of diagonal dominance in both matrices suggests an effective model. After sorting, diagonal values increased, suggesting enhanced performance within each class. Organizing tweets chronologically could improve the RF model’s precision in categorizing tweets.

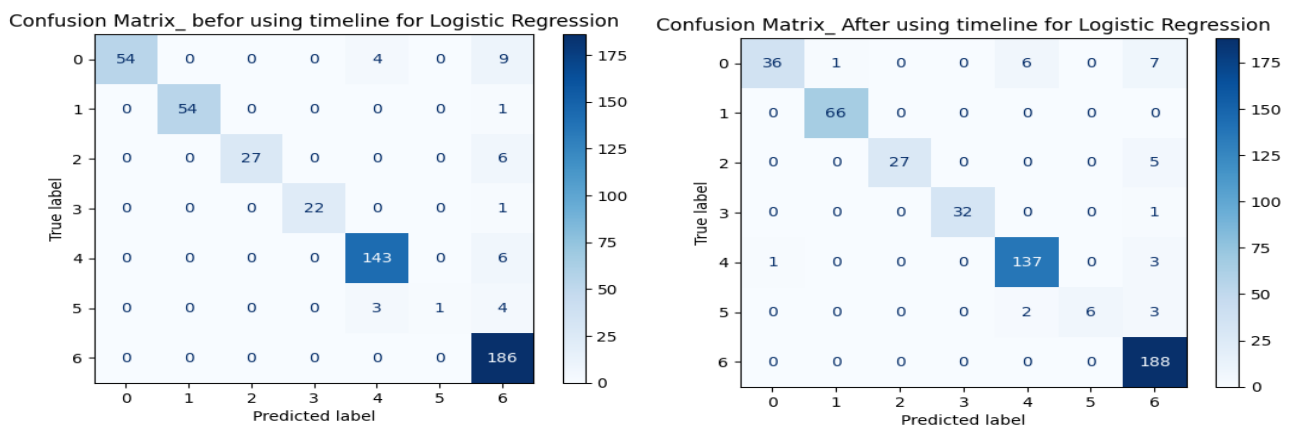


Fig. 4. Confusion matrix for Arabic tweet crime prediction for logistic regression.

The result in Fig. 4 for LR compares the model’s performance before and after sorting the dataset by timeline. An analysis of the observed changes’ confusion matrix shows that the initial matrix displayed accurate predictions for labels 4 and 6, but also included misclassifications. The second matrix displayed enhanced predictions for labels 0 and 1 based on chronology data, a slight reduction for 4, and increased accuracy for label 6. These matrices aid in visualizing the performance of prediction models and comprehending the influence of other characteristics, such as timeline data.

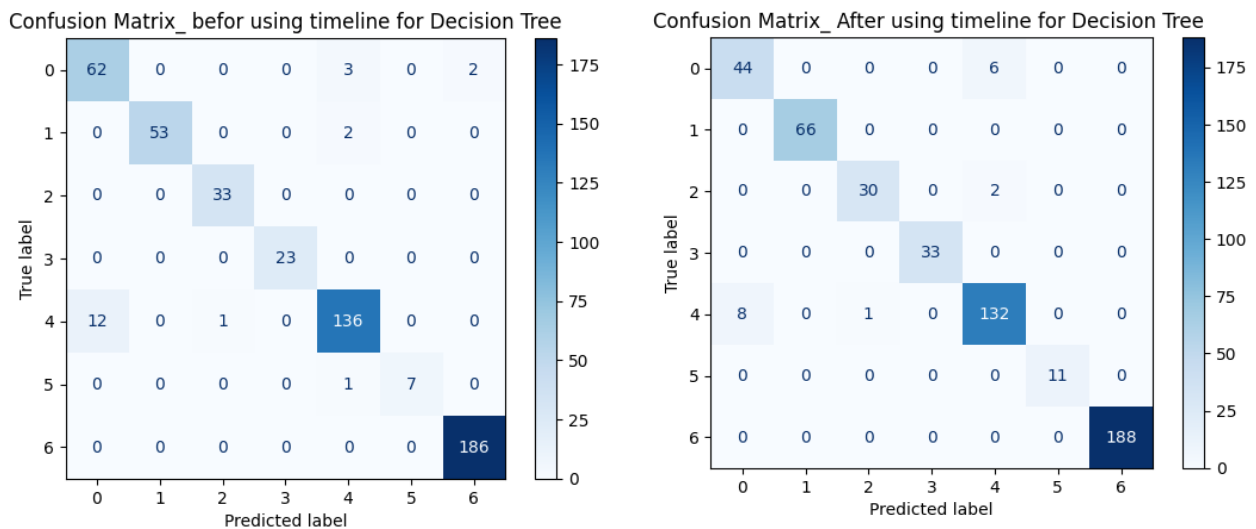


Fig. 5. Confusion matrix for Arabic tweet crime prediction for decision tree.

The result in Fig.5 for DT compares the model’s performance before and after sorting the dataset by timeline. The observed changes in the confusion matrix were analyzed. Before timing data were included, the model accurately predicted labels to a satisfactory degree but struggled to differentiate between labels 4 and 6. Utilizing chronology data

resulted in a minor loss in accuracy for label 0, while accuracy improved for labels 1 and 3. The performance results indicated a convergence between the values of RF and DT. The best model was selected by adding other metrics for examination, namely, AUC, mAP, and MCC, as shown in Table III.

TABLE III. PERFORMANCE RESULTS FOR PREDICTING CRIMES IN ARABIC TWEETS USING DIFFERENT METRICS.

| Model for the Proposed Method | AUC | MAP | MCC |
|---|---------------|---------------|--------------|
| Before sorting dataset data time RF + TF- IDF | 99.6% | 89.07% | 0.79% |
| After sorting dataset data time RF + TF- IDF | 99.57% | 94.04% | 0.16% |
| Before sorting dataset data time LR +TF- IDF | 99.79% | 83.70% | 0.16% |
| After sorting dataset data time LR +TF- IDF | 99.90% | 90.51% | 0.16% |
| Before sorting dataset data time DT +TF- IDF | 97.23% | 92% | 3.12% |
| After sorting dataset data time DT +TF- IDF | 97.99% | 93.44% | 3.57% |

When comparing the results and analyzing the metrics, the imbalanced nature of the dataset should be considered. Natural RF and DT exhibited differences when using an imbalance dataset to filter Arabic crime tweets. The best model is selected after using AUC, MAP, and MCC. A comparison of the findings and an analysis of the dataset show that RF has a higher AUC of 99.57% than DT does, which is 97.99%. A high AUC shows good positive/negative class discrimination by the model. The mAP, which evaluates the quality of the results, for RF and DT is 94.04% and 93.44%, respectively. With regard to MCC, RF has a very low MCC of 0.16%, while DT has a substantially higher MAP of 3.57%. A higher MCC indicates better overall performance, particularly in handling imbalanced data. Considering these metrics, while both models have similar performance in terms of precision, recall, F1-score, and accuracy, DT outperforms RF in terms of MCC. Therefore, on the basis of the provided metrics and considering the dataset’s imbalance, DT might be considered the better model based on the MCC metric. An analysis of relevant literature is shown in Table IV.

TABLE IV ANALYSIS OF RELEVANT LITERATURE ABOUT THE PROPOSED WORK.

| Paper authors | Filter crimes | The algorithm used to build the model | Accuracy and MCC |
|-----------------------------|--|--|--|
| [10] 2020 | Without the crime behavior filter | Brown clustering | - |
| [11]2020 | Without the crime behavior filter | Naïve Bayesian, RF, J48, and ZeroR | RF has the highest accuracy (98.1%), and ZeroR has the lowest accuracy (61.5%). |
| [12]2022 | Without the crime behavior filter | NB, DT, RF, LR, KNN, and SVM | 86%, 91%, 93%, 91%, 80%, and 93% |
| [13]2021 | Without the crime behavior filter | MNB and KNN or SVM | - |
| [14]2022 | Without the crime behavior filter | Lexicon-based and deep learning, with BERT | 94.91% |
| [15]2020 | Without the crime behavior filter | KNN and SVM | 89.4% and 92.4 |
| [16]2022 | Without the crime behavior filter | NB, LR, and Rf | 0.844, 0.859, and 0.866, respectively |
| [17]20 | Without the crime behavior filter | Improved RF, SVM, and naïve Bayesian model | 0.895, 0.848, and 0.809, respectively |
| Proposed method 2024 | An intelligent dictionary depends on a genetic algorithm that filters Arabic tweet crimes. | Prediction of Arabic tweet crimes using TF-IDF with DT, RF, and LR based on the timeline | 97% for DT, 97% for RF, and 94.43% for LR. DT is the best model based on an MCC of 3.57%. |

5. Conclusions

The study uses Twitter to analyze and predict Arabic-related crimes by using an intelligent dictionary based on a genetic algorithm for filtering abnormal behavior. It provides insights into user behavior related to criminal activities. As the first stage [19], and continues to complete the proposed method to predict Arabic-related crimes using ML. Given the results described in the previous section and due to the imbalanced nature of the datasets, evaluating performance based solely on metrics such as accuracy can be unclear and misleading. Therefore, we used other metrics that support this same type of dataset, such as MCC, which takes into account true positives, true negatives, false positives, and false negatives. Tests indicate that DT was the best of the three models. The findings can help governments and law enforcement in combating criminals. Future research should explore other social media platforms, ML algorithms, image analysis, and other locations and time frames for comparison. Our research offers valuable insights into the impact of ML methods on crime prediction. Nevertheless, there are still obstacles to overcome, such as the requirement for models that can be easily understood, explanations of causality, and precise incorporation of data into predictive models. Different places and periods can be used to enable comparisons of future outcomes by utilizing negative handling in ML for prediction.

Conflicts of Interest

The author declares no conflict of interest.

Acknowledgment

The authors would like to thank the Informatics Institute of Postgraduate Studies, Iraqi Commission for Computer & Informatics (<https://iips.edu.iq/>), Baghdad, Iraq, for their support in this work.

References

- [1] Bal Krishna Shah, Nitu Sharma, Saloni Bandgar and Prof. Sainath Patil, "Cybercrime Prevention on Social Media," *International Journal of Engineering Research & Technology (IJERT)*, Vols. Vol. 10 ,NO 03, March, no. ISSN: 2278-0181, 2021. DOI : [10.17577/IJERTV10IS030277](https://doi.org/10.17577/IJERTV10IS030277).
- [2] Brett Drury c d, Samuel Morais Drury e, Md Arafatur Rahman f, Ihsan Ullah, "A social network of crime: A review of the use of social networks for crime and the detection of crime", *Online Social Networks and Media*, vol. Volume 30, July 2022. DOI: [10.1007/978-981-15-5856-6_44](https://doi.org/10.1007/978-981-15-5856-6_44).
- [3] Jolan Rokan Naif¹, Ghassan H. Abdul-majeed², Alaa K. Farhan³, "Internet of Things Security using New Chaotic System and Lightweight AES," *AL-Qadisiyah for computer science and mathematics*, vol. Vol.11., No.2, 2019. DOI:[10.29304/jqcm.2019.11.2.571](https://doi.org/10.29304/jqcm.2019.11.2.571)
- [4] Z. Abbass, Z. Ali, M. Ali, B. Akbar, and A. Saleem, "A Framework to Predict Social Crime through Twitter Tweets By Using Machine Learning," *IEEE 14th International Conference on Semantic Computing (ICSC)*, pp. 363-368, 2020. DOI: [10.1109/ICSC.2020.00073](https://doi.org/10.1109/ICSC.2020.00073).
- [5] J. Pereira-Kohatsu, J.C.; Quijano-Sánchez, L.; Liberatore, F.; Camacho-Collados, M. Detecting and Monitoring Hate Speech in Twitter. *Sensors* **2019**, *19*, 4654. <https://doi.org/10.3390/s19214654>.
- [6] Yagcioglu, S.; Seyfioglu, M.S.; Citamak, B.; Bardak, B.; Guldamlasioglu, S.; Yuksel, A.; Tatli, E.I." "Detecting Cybersecurity Events from Noisy Short Text," *North American Chapter of the Association for Computational Linguistics (NAACL)*, p. 1366–1372, 2019. <https://doi.org/10.48550/arXiv.1904.05054>.
- [7] Fang, Y.; Gao, J.; Liu, Z.; Huang, C. Detecting Cyber Threat Event from Twitter Using IDCNN and BiLSTM. *Appl. Sci.* **2020**, *10*, 5922. <https://doi.org/10.3390/app10175922>.
- [8] I. Idrissi, M. Boukabous, M. Azizi, O. Moussaoui, and H. El Fadili "Toward a deep learning-based intrusion detection system for IoT against botnet attacks," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. vol. 10, no. 1, pp. 110-120,, Mar. 2021. DOI: [10.11591/ijeecs.v23.i2.pp1059-1067](https://doi.org/10.11591/ijeecs.v23.i2.pp1059-1067).

- [9] Abdalrda, Z.K., Al-Bakry, A.M., Farhan, A.K. (2023). A hybrid CNN-LSTM and XGBoost approach for crime detection in tweets using an intelligent dictionary. *Revue d'Intelligence Artificielle*, Vol. 37, No. 6, pp. 1651-1661. <https://doi.org/10.18280/ria.370630>.
- [10] Vo, Thanh, et al. 'Crime Rate Detection Using Social Media of Different Crime Locations and Twitter Part-of-speech Tagger with Brown Clustering'. 1 Jan. 2020 : 4287 – 4299. DOI: [10.3233/JIFS-190870](https://doi.org/10.3233/JIFS-190870).
- [11] V. Vijendra Singh, Vijayan K Asari, Kuan-Ching Li, "Analysis and Classification of Crime Tweets," *Procedia Computer Science*, vol. 167, pp. 1-2662, 2020. <https://doi.org/10.1016/j.procs.2020.03.211>.
- [12] Sheila Marie M. Matias; Jefferson A. Costales; Christian M. De Los Santos, "A Framework for Cybercrime Prediction on Twitter Tweets Using Text-Based Machine Learning Algorithm," in 2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI), Chengdu, China, 19-21 August 2022. DOI: [10.1109/PRAI55851.2022.9904212](https://doi.org/10.1109/PRAI55851.2022.9904212).
- [13] Sreya C M , "A Framework To Predict Social Crimes Using Twitter Tweets," (IJRASET) *International Journal for Research in Applied Science & Engineering Technology*, vol. 9, no. 1, Jan 2021. <https://doi.org/10.22214/ijraset.2021..32879>.
- [14] Mohammed Boukabous, Mostafa Azizi, "Crime prediction using a hybrid sentiment analysis approach based on the bidirectional encoder representations from transformers," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 25, no. 2, pp. 1131~1139 ISSN: 2502-4752, February 2022. DOI: [10.11591/ijeecs.v25.i2.pp1131-1139](https://doi.org/10.11591/ijeecs.v25.i2.pp1131-1139).
- [15] Aditi Sarker, Partha Chakraborty, S. M. Shaheen Sha, Mahmuda Khatun, Md. Rakib Hasan, Kawshik Banerjee, "Improvised Technique for Analyzing Data and Detecting Terrorist Attack Using Machine Learning Approach Based on Twitter Data," *Journal of Computer and Communications*, vol. 8, no. 7, July 30, 2020. DOI: [10.4236/jcc.2020.87005](https://doi.org/10.4236/jcc.2020.87005).
- [16] Aljwari, F., Alkaber, W., Alshutayri, A., Aldahri, E., Aljojo, N., & Abouola, O. (2022). Multi-scale Machine Learning Prediction of the Spread of Arabic Online Fake News. *Postmodern Openings*, 13(1 Sup1), 01-14. <https://doi.org/10.18662/po/13.1Sup1/411>
- [17] Wotaifi, Tahseen A. and Dhannoon, Ban N. (2022) "Improving Prediction of Arabic Fake News Using Fuzzy Logic and Modified Random Forest Model," *Karbala International Journal of Modern Science: Vol. 8 : Iss. 3 , Article 18*. Available at: <https://doi.org/10.33640/2405-609X.3241>
- [18] Twitter Developer. (2021), <https://developer.twitter.com/en/docs/twitter->.
- [19] Zainab Khyioon Abdalrda , Abbas Mohsin Al-Bakry and Alaa K. Farhan, "Analysis of social networks and filtering of Arabic crime tweets based on an intelligent dictionary using a genetic algorithm", *Global Journal of Engineering and Technology Advances*, 2024, 18(02), 177–191. <https://doi.org/10.30574/gjeta.2024.18.2.0033>.
- [20] Samah M. Alzanin, Aqil M. Azmi, Hatim A. Aboalsamh, Short text classification for Arabic social media tweets," *Journal of King Saud University - Computer and Information Sciences*. 34(9), 6595-6604. (2022). <https://doi.org/10.1016/j.jksuci.2022.03.020>.
- [21] Sarker, I.H.; Abushark, Y.B.; Alsolami, F.; Khan, A.I. IntraDTree: A Machine Learning Based Cyber Security Intrusion Detection Model. *Symmetry* **2020**, *12*,754. <https://doi.org/10.3390/sym12050754>.
- [22] Hussain Ali, Y., Sabu Chooralil, V., Balasubramanian, K., Manyam, R.R., Kidambi Raju, S., T. Sadiq, A., Farhan, A.K. (2023). Optimization system based on convolutional neural network and internet of medical things for early diagnosis of lung cancer. *Bioengineering*, 10(3): 320. <https://doi.org/10.3390/bioengineering10030320>.
- [23] Jaleel, H.Q., Stephan, J.J., Naji, S.A. (2022). Textual dataset classification using supervised machine learning techniques. *Engineering and Technology Journal*, 40(04): 527-538. <https://doi.org/10.30684/etj.v40i4.1970>.
- [24] Dhall, D., Kaur, R. and Juneja, M., 2020. Machine learning: a review of the algorithms and its applications. *Proceedings of ICRIC 2019*, pp.47-63. https://doi.org/10.1007/978-3-030-29407-6_5.



-
- [25] Zhang, H., Xiao, X., Mercaldo, F., Ni, S., Martinelli, F. and Sangaiah, A.K., 2018. Classification of ransomware families with machine learning based on N-gram of opcodes. *Future Generation Computer Systems*, 90, pp.211-221. DOI: [10.1016/j.future.2018.07.052](https://doi.org/10.1016/j.future.2018.07.052).
- [26] Rifai, Hozayfa El, Leen Al Qadi, and Ashraf Elnagar. "Arabic Multi-label Text Classification of News Articles." In *Advanced Machine Learning Technologies and Applications: Proceedings of AMLTA 2021*, pp. 431-444. Springer International Publishing, 2021. DOI: [10.1007/978-3-030-69717-4_41](https://doi.org/10.1007/978-3-030-69717-4_41).
- [27] Al Hamad, Mona, and Ahmed M. Zeki. "Accuracy vs. cost in decision trees: A survey." In *2018 international conference on innovation and intelligence for informatics, computing, and technologies (3ICT)*, pp. 1-4. IEEE, 2018. DOI: [10.1109/3ICT.2018.8855780](https://doi.org/10.1109/3ICT.2018.8855780).
- [28] Y. Dhebar and K. Deb, "Interpretable rule discovery through bilevel optimization of split-rules of nonlinear decision trees for classification problems," *IEEE Trans. Cybern.*, vol. 51, no. 11, pp. 5573–5584, 2021, DOI: [10.1109/TCYB.2020.3033003](https://doi.org/10.1109/TCYB.2020.3033003).
- [29] Charbuty, Bahzad, and Adnan Abdulazeez. "Classification based on decision tree algorithm for machine learning." *Journal of Applied Science and Technology Trends* 2, no. 01 (2021): 20-28. DOI: <https://doi.org/10.38094/jastt20165>.
- [30] Zhang, Z., Sabuncu, M. (2018). Generalized cross-entropy loss for training deep neural networks with noisy labels. *Advances in Neural Information Processing Systems*, 31: 8778–8788. <https://doi.org/10.48550/arXiv.1805.07836>.
- [31] Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*. 2020 Jan 2;21(1):6. DOI: [10.1186/s12864-019-6413-7](https://doi.org/10.1186/s12864-019-6413-7).
- [32] Arya TafviziBesim AvciMukund Sundararajan, "Attributing AUC-ROC to Analyze Binary Classifier Performance", (2022). *Attributing AUC-ROC to Analyze Binary Classifier Performance*. 10.48550/arXiv.2205.11781.
- [33] Wang, Beinan. (2022). *A Parallel Implementation of Computing Mean Average Precision*. 10.48550/arXiv.2206.09504.