Revue d'Intelligence Artificielle
Vol. 37, No. 6, December, 2023, pp. 1651-1661
Journal homepage: http://iieta.org/journals/ria

A Hybrid CNN-LSTM and XGBoost Approach for Crime Detection in Tweets Using an Intelligent Dictionary

Zainab Khyioon Abdalrdha[1*], Abbas Mohsin Al-Bakry[2], Alaa K. Farhan[3]

[1] Iraqi Commission for Computers and Informatics / Informatics Institute of Postgraduate Studies, Baghdad 10001, Iraq
[2] University of Information Technology and Communication (UoITC), Baghdad 10001, Iraq
[3] Department of Computer Sciences University of Technology, Baghdad 10001, Iraq

Corresponding Author Email: phd202120695@iips.edu.iq

https://doi.org/10.18280/ria.370630

Received: 26 August 2023
Revised: 1 October 2023
Accepted: 7 October 2023
Available online: 27 December 2023

Keywords:
crime detection, machine learning, intelligent dictionary, text classification, graph analysis, natural language processing (NLP), deep learning

ABSTRACT

As social media grows, recognizing and managing illicit content, including threats, harassment, hate speech, armed robbery, drug smuggling, blackmail, and other crimes, is crucial. The present study uses machine learning and deep learning to create an intelligent lexicon for identifying crime-related material in Twitter tweets. The Aho-Corasick technique effectively creates a dictionary for extensive text corpus search, categorization, and keyword-based action execution. This strategy overcomes the evolution of language dynamics and criminal vocabulary to improve crime-related information detection. This paper aims to gather accurate data to help law enforcement identify and prevent specific crimes. For tweet preprocessing and extraction of relevant information like textual patterns and other distinctive qualities, natural language processing (NLP) technologies are prioritized. The paper describes labeling tweets into crime categories. This dataset trains supervised learning models to categorize tweets as criminal or not. XGBoost and Hybrid CNN-LSTM are combined for this. The suggested technique is assessed using precision, recall, F1-Score, accuracy, and MAP accuracy measures. These metrics measure the model's crime-related tweet identification accuracy. The Arabic tweets dataset, encompassing 18493 tweets and 10 features, is utilized for model testing. After training, the Hybrid CNN-LSTM model demonstrated an accuracy of 99.84% and a macro F1-Score of 98.20%. When the XGBoost method was employed, the traditional machine learning model achieved a peak F1 macro score of 99.36% and a maximum accuracy of 100%. The results suggest that while the deep learning models outperform machine learning models in the F1-Score, XGBoost exhibits superior accuracy. The paper presents a comprehensive strategy for crime detection in tweets, potentially offering a significant tool for law enforcement agencies.

1. INTRODUCTION

The fact that crime rates have increased in Iraq and the Arab World in general is a key step toward resolving a problem addressed in this paper. Crime rates have risen as a result, including electronic extortion, theft, bullying, narcotics, and a slew of other offenses. Understanding the growth and regionalization of crime requires an examination of crime statistics. Law enforcement faces challenges as a result of the amount of information available, technological improvements, and population density. The current study was motivated in large part by the growth in crime rates and the spread of these crimes through social networking sites and the study's goal is to collect trustworthy information to assist in characterizing the nature of crime and enhance crime prevention. Because of the complexities of the Arabic language, it is more difficult to detect crimes, which boosts human security. Crime impacts society through social media, with bullies targeting individuals based on race, body image, or religious beliefs [1]. Crime impacts individuals, society, and finances, requiring significant efforts for prevention, victim care, and legal proceedings to improve lives and security [2]. Social networks facilitate cultural, political, and historical insights and knowledge sharing [3]. Social media facilitates crimes and the different forms these crimes take, particularly through social engineering assaults, where users are tricked into providing personal information or clicking on phishing links. Social media platforms can be used by criminals to spread malware through links and attachments, create phony accounts, publish harmful websites, and distribute malicious files.

Cybercriminals can use social media data to commit identity theft, impersonate individuals, or steal their identities to carry out fraudulent activities [4]. The era of rapid ICT development has influenced modern society, altering space, time, and identity relationships [5]. Rapid ICT development has changed geography, time, and identity interactions in modern society. Crime has been significantly impacted by the development of information and communication technologies, or ICTs. It has increased the probability of criminal activity, including fraud and identity theft, as well as the sophistication of criminal

activity. Additionally, ICTs have made it simpler for thieves to execute their plans quickly and effectively. They have, however, made it more challenging for law enforcement to identify and stop crime as they may use encryption and other technologies to hide their activities. Due to the rise in cybercrime, law enforcement must stay up to date on new developments in technology and adopt innovative tactics [6]. Crime in cultures impacts individuals and communities, requiring studies to understand behavior, identify individuals, and effectively detect, anticipate, and prosecute [7]. Governments are implementing various strategies to prevent crime on social media. These include promoting crime prevention through social media, conducting social media surveillance, monitoring criminal activity, and regulating social media companies. These measures aim to address privacy and civil liberties concerns, gather valuable information about crimes in progress, and hold social media companies accountable for illegal content [8]. Social media technology revolutionizes interactions through platforms like Facebook, blogs, wikis, and Twitter [9]. This paper proposes a technique using machine learning and deep learning models to classify keywords in crime tweet datasets. The dataset evaluates performance metrics and compares results. Improving convolutional neural networks (CNN) using LSTM after improving becomes Hybrid CNN-LSTM is employed to classify new tweets as crime-related or non-crime-related. Machine learning and deep learning methodologies are well-suited for the task of crime detection on social media platforms due to their capacity to effectively evaluate vast quantities of data and discern intricate patterns that may prove challenging for human observers to discern. Furthermore, these systems can acquire knowledge from novel data and enhance their precision as time progresses. Moreover, deep learning algorithms possess the capability to effectively handle unstructured data, including photos and text, thereby playing a crucial role in the analysis of social media material.

The following are the work's main contributions:

1. Creating a new dataset in the language Arabic tweets crimes that have been enhanced to include more features and enough data that is adequate for these types of crimes using a Python library to scrape tweets and the Aho-Corasick algorithm to extract tweets with approximately 38 keywords related to crimes and approximately 25 keywords unrelated to crimes.

2. This is the first thorough study that compares DL and ML models for Arabic criminal tweet identification tasks for various keywords, evaluating their performance in Arabic crime tweet detection tasks.

This paper is structured as follows: The related work is presented in Section 2; next, Section 3 describes how to create an intelligent dictionary to gather the dataset; next, Section 4 describes the intelligence tools that were used to create the proposed model; next, Section 5 describes the proposed model's execution; and finally, Section 6 presents the results and discusses them. Finally, in Section 7, conclusions are reached and potential directions for further study are discussed.

2. RELATED WORK

Social media has evolved into a powerful communication tool for sharing news and expressing feelings on a variety of themes, particularly in Arabic, a difficult language with fewer resources than English and Chinese in addition, Arabic language instruction is popular, local information is easily accessible, and cultural idiosyncrasies make it a good tool for spotting illegal activities on Twitter. Dialects, colloquialisms, false positives, the amount of data, and ethics are some of the challenges. For responsible and accurate Arabic Twitter Crime Detection, linguistic competency, natural language processing, and cultural awareness are required. Instead of focusing on a single crime, this study looked at multiple crimes [10]. The study examines Twitter crimes in Arabic using DL and ML methodologies, comparing English outcomes. Machine learning (ML) and deep learning (DL) models are well-suited for the analysis of social media material owing to their capacity to process substantial volumes of data and discern intricate patterns. Deep learning algorithms can gradually acquire high-level features, enhance accuracy through iterative processes, and obtain advanced characteristics through the practice of feature engineering. Convolutional neural networks (CNNs) have demonstrated superior performance compared to conventional machine learning methods across several domains, including but not limited to cybersecurity, natural language processing, bioinformatics, robotics, and medical information processing. Deep learning algorithms possess the capability to acquire knowledge from extensive datasets and exhibit expedited testing durations, rendering them well-suited for real-time applications [11]. Kaddoura et al. [12] offer a strategy for identifying Arabic tweets using machine learning and deep learning techniques, using feature extraction and N-gram models, and computing performance metrics. With an F1-Score of 99.73%, the neural network method outscored the other algorithms, but GloVe outperformed rapid Text by 0.5%. Support vector machines, neural networks, logistics regression, and naive Bayes are the classical machine-learning techniques used. Deep learning approaches make use of global vector (GloVe) and fast Text learning models. The study [13] used DL models to identify offensive remarks on Arab YouTube channels, including Bi-LSTM with attention mechanism, CNN, Bi-LSTM, and CNN-LSTM. Results showed accuracy of 85.7%, 86.4%, 87.2%, and 87.8%, with CNN outperforming competitors. CNN, and CNN-LSTM models, were employed by the authors in this paper [14], in their "fast and simple" methods. On a dataset of 8K Arabic tweets, they tested each model. According to the findings, ML models cannot compete with neural learning models. Using the hybrid model CNN-LSTM, the best F1-Score was 73%. This paper's authors [15], The text analysis algorithm categorizes tweets as cyberbullying using TF-IDF and Word2Vec, using seven machine learning classifiers. The model achieves a 96.4% F1-Score rate and outperforms other entity recognition methods [16]. used ML algorithms like NB, DT, RF, and SVM to classify religious hate speech in Arabic tweets. They trained DL models using Fast Text and word2vec embedding, achieving a high F-Measure and accuracy of 71%. Finally, Aljarah et al. [17], used machine learning and natural language processing to identify cyber hate speech on Twitter in Arab environments, with the RF method achieving 91.3% accuracy.

3. BUILD AN INTELLIGENT DICTIONARY TO COLLECT THE DATASET

There are other ways to collect datasets from Twitter, including using the official Twitter API offered by the Twitter development team and developers can programmatically

access Twitter's data via the Twitter API. It lets users access tweets, user data, trends, and Twitter activities. Developers require a Twitter Developer Account and application to get API keys and access tokens. Twitter's API endpoints include GET /tweets/search/recent, GET /users/show, GET /followers/list, and POST /update. Developers must limit API request rates to prevent abuse. The returned JSON data can be saved in databases or data lakes for analysis. Developers may monitor events and trends with real-time streaming APIs for tweets and data [18], this paper will discuss how to build a dictionary in this section. By utilizing the numerous library tools offered by the Python programming language, such as Tweepy and Scrape. The Aho-Corasick algorithm was used by the "sn scrape" library to Scrap the dataset.

3.1 Twitter data acquisition

This study collects Twitter data to assess trained classifiers' performance in real-world scenarios, utilizing up-to-date tweets from accounts. Experiments are designed to ensure practical applicability and data collection for effective detection.

3.2 Twitter data extraction

This study collects Arabic tweets using the Python library, using keywords for crime and non-crime content. The Aho-Corasick algorithm is used to build an Intelligent method that can be used for graph and metadata analysis, classifying tweets related to crime. Recent tweets are extracted for analysis before building models, focusing on detecting crimes in Arabic tweets. Some of these keywords that were used are shown in Table 1. The keywords used in building the Intelligent Dictionary.

Table 1. The keywords used in building the intelligent dictionary

No. keywords	Keywords Arabic	Keywords English
1	إرهاب	terrorism
2	'تنظيم داعش	ISIS
3	"تنمر'	bullying
4	المخدرات	drugs
5	الممنوعات	contraband
6	'لاحتيال الالكتروني	electronic fraud
7	التزوير الالكتروني	electronic forgery
8	سرقة	theft
9	سطو	burglary
10	سرقة مسلحة	armed robbery
11	سرقة بنك	bank robbery
12	تهريب	smuggling

(Continue **Table 1**)

No. keywords	Keywords Arabic	Keywords English
13	تهريب المخدرات	drug smuggling
14	تهريب البشر	human smuggling
15	تهريب السلاح	arms smuggling
16	قتل	murder
17	'قتل غير متعمد'	manslaughter
18	اختطاف	kidnapping
19	'اختطاف الأطفال	children kidnapping
20	'اختطاف للحصول على فدية	kidnapping for ransom
21	ابتزاز	blackmail
22	الابتزاز الإلكتروني	electronic blackmail
23	تزوير	forgery
24	تزوير توقيع	forging a signature
25	تزوير وثائق	forging documents
26	تزوير جواز سفر	passport fraud
27	غسيل الأموال	money laundering
28	اختطاف	Kidnapping
29	تمويل الإرهاب	terrorist financing
30	جريمة منظمة	organized crime
31	'تجارة البشر	human trafficking
32	'الاختراق الإلكتروني	electronic hack
33	تجسس	to spy
34	التجسس الإلكتروني	electronic espionage
35	تجسس صناعي	industrial espionage
36	مخترق	penetrative
37	هجمات إلكترونية'	cyberattacks
38	الكوكايين	cocaine
39	حقوق الإنسان	human rights
40	السلم	the ladder
41	الأمن	Security
42	التعاون	cooperation
43	التضامن	solidarity
44	العدالة والمساواة	justice and equality
45	الإيجابية	positive
46	التحفيز	Stimulus
47	التسامح والاحترام	tolerance and respect
48	التعليم والثقافة	education and culture
49	العمل الخيري والتطوع	charitable work and volunteering
50	الإلهام	inspiration
51	التحفيز	stimulus
52	الإنجاز	achievement
53	التفاؤل	optimism
54	التطور	development
55	التأمل	meditation
56	التجدد	regeneration
57	التقدم	progress
58	التناغم	harmony
59	السرور	pleasure
60	النشاط	activity
61	السعادة	happiness
62	التأمل	Meditation
63	الانتعاش	recovery

Table 2. Feature extraction from tweets

No.	Feature Linked to the Tweet	Description
1	Tweet Id	A unique tweet ID is assigned to each tweet.
2	Username	Twitter users' unique usernames, identifying their account on Twitter.
3	Tweet	The user's tweet's text content includes messages, information, or statements in limited characters.
4	Quote Tweets	Tweet's number of quotes from other users.
5	Likes	It represents the number of likes or favorites received by the tweet.
6	Retweets	Retweet count indicates how many times a tweet has been retweeted.
7	Country	It represents the location or country associated with the user who posted the tweet.
8	Date time	Date and time of tweet creation.
9	Interactions	It represents the overall number of interactions or engagements with the tweet, which can include likes, retweets, replies, and other forms of engagement.

The data collection process targeted various fields, including politics and journalism. The classification process used criminal tweets, with a total of 1,19597 tweets. Where the total number of tweets (1,19597), the number of positive tweets (60524), the number of similar tweets (50427), the number of bad tweets about criminal activity (67011), and the number of identical tweets (55254). The data set was divided into six parts, and 18,493 data sets were taken for training and testing in this paper. It relied on the Aho-Corasick algorithm to pull tweets from Twitter and, with the help of Python tools linked with Twitter within scraping, collected data that contained crimes and those that were free. Table 2 shows the most important feature of extracting from tweets.

4. INTELLIGENT TOOLS USED IN BUILDING THE PROPOSED MODEL

The main smart tools used to develop the proposed approach are described in this section.

4.1 Aho-Corasick algorithm

Developed in 1975 by Aho and Corasick, the Aho-Corasick algorithm compares input text patterns using a tree-like structure, using a finite state machine and common prefixes [19]. The Aho-Corasick algorithm is an efficient string-matching method using a tree structure and failure transitions. It is widely used in applications like intrusion detection, virus scanning, text mining, and lexical due to its linear time complexity. The preprocessing step builds the automaton once, allowing multiple patterns to be compared against input texts. The utilization of the Aho-Corasick algorithm in constructing a dictionary can be characterized as advantageous in situations when there is a requirement to efficiently and effectively search a substantial collection of terms within texts. After identifying the keywords inside the text, one can categorize them or execute a certain action by the identified term. The construction of the dictionary is designed to align with the proposed approach and employ outcomes to identify criminal activities associated with Twitter posts.

4.2 Natural language processing (NLP) technique

This section demonstrates how to efficiently operate machine learning algorithms by representing text documents as numerical vectors and using the TF-IDF technique to extract features in NLP tasks. Because TF-IDF can recognize significant words in a document or corpus and give them weights depending on their significance, it is useful for feature extraction. It can handle big datasets: TF-IDF is scalable and capable of handling big datasets, which makes it appropriate for processing massive volumes of text data.

4.3 Basic machine learning

Computer scientists have focused on machine learning since the 1950s [20]. ML applications include translation and sentiment analysis [21]. ML algorithms are categorized into supervised, semi-supervised, and unsupervised types based on application goals [22, 23]. Supervised algorithms train using labeled data for tasks like crime detection and sentiment classification, using techniques like SVM, Naive Bayes, Decision Trees, Random Forests, KNN, and Logistic Regression [24, 25]. Semi-supervised algorithms use labeled and unlabeled data during training, which is useful when limited or expensive labeled data is scarce. Unsupervised algorithms, like K-Means and DBSCAN, focus on discovering patterns, clusters, or relationships in unknown, unlabeled data [26]. ML algorithms revolutionize tasks like translation and sentiment analysis, advancing techniques and applications [21]. Understanding the history of machine learning can help us appreciate XGBoost, a popular ensemble learning algorithm family [27]. It's beneficial for crime detection as it transitions from rule-based AI to data-driven methods. XGBoost excels in processing and learning from big datasets, making it a valuable tool in crime detection-based intelligent Dictionaries for a method proposed.

4.4 Deep learning (DL)

Deep learning extracts intricate features from dimensional data, creating models connecting inputs to outputs using multi-layered networks and layered neural networks for abstract computation [28]. DL techniques improve accuracy and reduce training time in complex problems, enabling breakthroughs in science, engineering, and engineering tasks like data analysis, pattern recognition, and prediction [29]. DL algorithms excel in autonomous vehicles, robotics, intelligent systems, community-related applications, and social network analysis, uncovering patterns and trends for valuable insights into user behavior and preferences [30]. DL models offer potential in health, imaging, disease diagnosis, drug discovery, and personalized medicine. DL algorithms are supervised and unsupervised, requiring labeled training data for predictions [31-33]. Deep learning revolutionizes data analysis, pattern recognition, decision-making, and driving innovation. These deep learning models consist of two parts, A and B:

A. CNN (Convolutional neural network)

CNNs are used for image analysis and text processing in cybercrime detection. They use convolutional filters to capture local patterns and features, identifying keywords, linguistic patterns, and structural characteristics. For additional processing or categorization, the output is passed into connected layers [34]. Furthermore, CNN has a specific architecture that consists of numerous hidden layers and is a Multi-Layered Perceptron network (MLP) [35, 36].

B. LSTM (Long short-term memory)

Long short-term memory (LSTM) networks, a subtype of recurrent neural networks, are capable of storing enormous volumes of data in a short period [37]. LSTMs recognize long-term dependencies, sequential patterns, and temporal correlations in data, making them ideal for handling text, modeling context, and detecting cybercrime [38]. Because CNN and LSTM are both potent deep learning models that work well with sequence prediction problems including spatial inputs, such as text data, they were selected for this model. While the LSTM is employed for sequence modeling and prediction, the CNN is utilized for feature extraction, which is crucial for text classification tasks. Because the LSTM can selectively ignore irrelevant information and remember previous inputs, it is especially helpful for sequence prediction issues. These two models work together to forecast material linked to crime in social media posts and to extract significant

features from the text data. Furthermore, it has been demonstrated that CNN and LSTM perform better than conventional machine learning algorithms in a variety of fields, such as sequence prediction and text classification. To identify crime tweets on Twitter, a hybrid model combining these two algorithms was employed.

5. PROPOSED MODEL

The proposed methodology employs Twitter as a news-oriented platform for user engagement, enabling users to communicate through responses, likes, and comments on diverse material. The five parts of the proposed model are broken down into the following subsections for description:

1. Twitter Data Extraction involves topic-based query using Python libraries, saving tweet datasets in a CSV file, as described in Section (2.3).
2. Preprocessing steps in dataset preparation include tokenization, removing stop words, handling special characters, normalization, and normalization for Arabic tweets, including spelling variations and morphological structures.
3. Techniques for extracting features include both conventional ones like a bag of words, TF-IDF, and n-grams as well as more sophisticated ones like word embedding and contextual embedding. For feature extraction, this study uses TF-IDF.
4. Classification using machine learning based on XGboost and improving deep learning model using Hybrid CNN_LSTM Model depends on task complexity, data size, and computational resources.
5. The model training includes feature extraction and model performance training using the hybrid CNN _LSTM model and training 80% of the training dataset.
6. The proposed model was evaluated using confusion matrices, MAP accuracy, and unbiased test dataset comparison.

5.1 Dataset labeling process

Data collection and classification process based on criminal tweets, identifying crimes and non-crimes using a dictionary and algorithm, labeling 0 and 1.

5.2 Dataset preprocessing

Data preprocessing involves operations to classify and clean collected data. involves steps:

General Cleaning: Cleansed dataset by removing links, spaces, and unnecessary elements.

Normalization: The normalization process unifies Arabic letter forms, ensuring consistency and embedding.

Diacritic Removal: Arabic diacritics were removed to unify similar word forms. Diacritics such as "ـَ" (Fatha), "ـُ" (Damma), "ـِ" (Kasra), and "ـّ" (Shadda) were removed. This step aimed to handle different variations of the same word without diacritics.

Removal of Repeated Letters: Removed repeated letters within words. For example, words like "عاااا" and "ووو_ح" were transformed back to their original form, "ع" and "و_ح," respectively. However, if a word contained two similar characters (e.g., "تسلل" and "نوع-"), deleting the repeated characters would change the meaning of the word ("تسل منوع" and "نوع-م"). To address this, the decision was made to remove repeated characters that occurred more than twice.

Punctuation Removal: The dataset was cleaned by removing punctuation marks. Examples of removed punctuation marks include ("<؟ () *&^%]] - `÷× _؛":/،,'{} ~¦+|!” …“–"). The goal was to eliminate punctuation that might affect the analysis or modeling process.

Number and Non-Arabic Character Removal: Data preprocessed for Arabic text analysis, removing numbers and non-Arabic characters. Figure 1 shows an example of a tweet before and after preprocessing.

	Username	Tweet	Datetime	Country	Likes	Retweets	Replies	Quote Tweets	Interactions	label	Clean_Tweet
0	abunouh991	ارهاب فكري @Alibakkar78	2023-06-11 11:32:21+00:00		0	0	0	0	0	0	ارهاب فكري
1	abopakar1090	#قريب\n\n....تغييرات جذرية بالمحرر وجميع الف...	2023-06-11 11:23:29+00:00		0	0	0	0	0	0	قريبا تغييرات جذرية بالمحرر وجميع للفصائل ...الثو
2	mohmed_qr	@ya_milhdi الجيش مدرب وعنده أسلحة يكثر يدخل معاركه ...معا	2023-06-11 11:19:56+00:00	Iraq	0	0	1	0	1	0	الجيش مدرب وعنده أسلحة يكثر يدخل معاركه ...والمعارك
3	thatsavageheo	@LaythDrk الكرك حرب طائفية في العراك \n و...كم	2023-06-11 11:18:27+00:00	كوكب زمردة	0	0	0	0	0	0	الكرك حرب طائفية العراك منظمة إرهابية فقراء ...مش
4	m70101413	@MayaSabbagh7 للمعلومية هذا حكم القتل ... بجانبين	2023-06-11 11:13:59+00:00	Riyadh	0	0	0	0	0	0	للمعلومية حكم القتل بجانبين السنه تناسبهم ...حتي
...	...	...	...	...	...	...	...	...	...	...	...
119592	PerfumesMap	...انعش ص\n!! لأن الانتعاش والتفرّد قد يكمن في عطر	2023-02-19 05:10:26+00:00	Riyadh	5	1	0	0	6	1	لأن الانتعاش والتفرّد يكمن عطر انعش صباحك ...بنفح
119593	cillo_ff	صباح الانتعاش	2023-02-19 03:24:09+00:00	الاتنتس	0	0	0	0	0	1	الانتعاش
119594	AsharqBusiness	...لجنة تنظيم الأوراق المالية الصينية حققت حده ال	2023-02-19 02:00:00+00:00		4	2	1	0	7	1	لجنة تنظيم الأوراق المالية الصينية حققت حده ...ال
119595	RehabTariqq	صباح الانتعاش https://t.co/suN4OCSTxh	2023-02-19 01:11:42+00:00		2	1	4	0	7	1	الانتعاش
119596	3etrAlsharq	من الحضور التي\n أكوا نيجو: برقيوم الإنتعاش ...ت	2023-02-18 23:52:01+00:00	Egypt	0	0	0	0	0	1	...أكوا نيجو برقيوم الإنتعاش الحضور التي تقييمها

119597 rows × 11 columns

Figure 1. Example tweet before and after preprocessing

5.3 TF-IDF feature extraction

The input data are converted into a set of features at this stage via feature extraction [39]. TF-IDF technique extracts features in natural language processing, valuing word importance by weighting frequent, rare words [40]. The following equations (1, 2) display the TF-IDF formula.

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

$$TF - IDF\ (t, d) = TF\ (t, d) \times \log\ (N/\ (DF\ (t) + 1)) \qquad (2)$$

The proposed technique utilizes a pre-processed TF-IDF dataset for machine learning classifiers. TF-IDF is an easy-to-understand, effective, and efficient feature extraction method. It gives significant word weights depending on their importance, employs a statistical technique to assess the relevance of terms in the text, and produces findings that are easy to understand. It is simpler to comprehend and interpret because it works well with short datasets and keyword-based classification tasks. Compared to word embeddings and contextual embeddings, TF-IDF requires less computing power, which makes it a better option for feature extraction.

5.4 Model construction

The suggested model for detecting tweet crimes compares ML and improving DL models trained on a dataset—Figure 2. The block diagram of the suggested technique framework provides examples of the proposed models for Twitter Crime Detection. The scheme uses a machine learning model, such as XGBoost, with a TF-IDF vectorizer. In Deep learning, the model is trained on a crime tweet dataset. The hybrid CNN _LSTM model is improved after increasing the CNN layers' complexity as the first development method. The development mechanism for the Model is explained in section A.

A. Hybrid_CNN LSTM model

A total of 13 layers make up the model, including Conv1D, Embedding, Max Pooling, LSTM, Dense, and Output layers. Filters on the input layer number 16, those on the second layer number 32, and the kernel size is 3. The maximum pooling layer has a size of 2, the third and fourth layers each have 64 or 128 filters, and the fifth layer has 256 filters and a kernel size of 3. The dense layer has 64 units and a ReLU activation function, whereas the LSTM layer captures a long dependency range in serialized data, including 64 units. As shown in Table 3 under the heading Structure of the Hybrid CNN _ LSTM Model and in Figure 3 under the heading Architecture of the Hybrid CNN LSTM Model, the output layer generates a probability score for the binary classification of whether or not the Tweet comprises a crime.

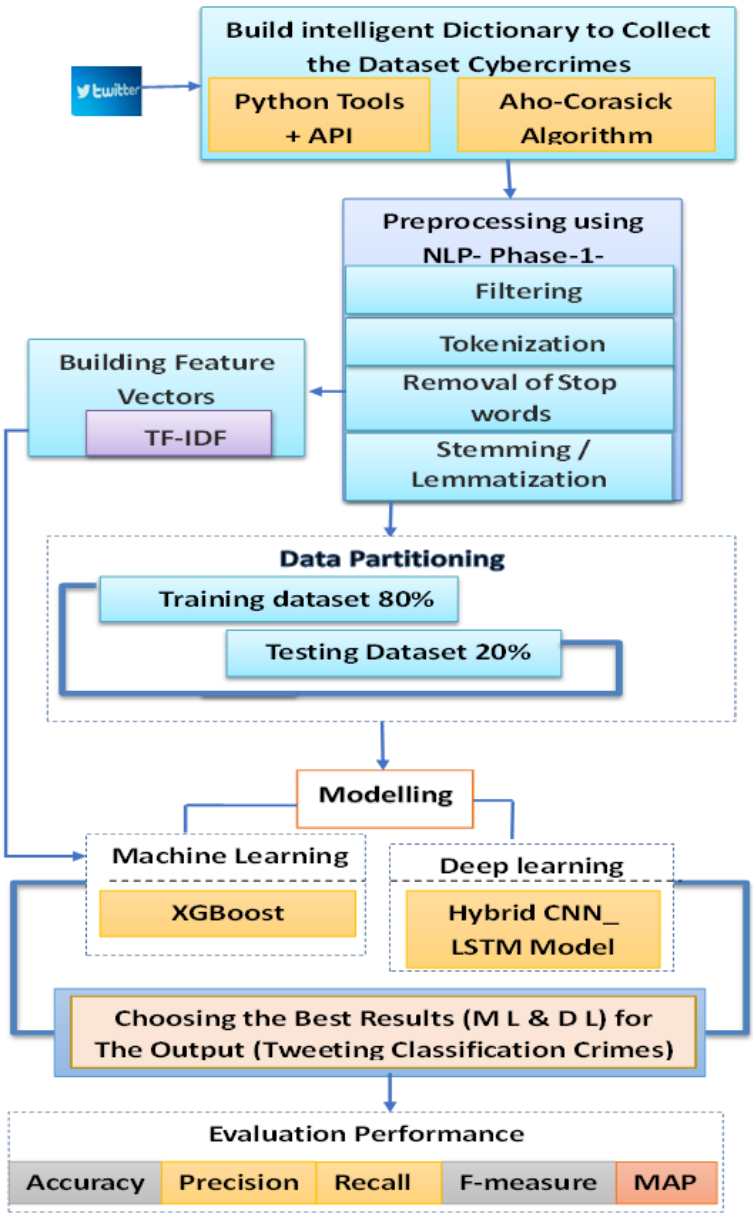

Figure 2. Block diagram of the suggested technique framework

Table 3. Hybrid_CNN LSTM model's structure

Layer (Type)	Output Shape	Parameter
conv1d_63 (Conv1D	(None, 44, 16)	64
conv1d_64 (Conv1D	(None, 44, 32)	1568
max_pooling1d_36 (Max Pooling1D	(None, 22, 32)	0
conv1d_65 (Conv1D)	(None, 22, 64)	6208
conv1d_66 (Conv1D)	(None, 22, 128)	24704
max_pooling1d_37 (Max Pooling1D)	(None, 11, 128)	0
conv1d_67 (Conv1D)	(None, 11, 256)	98560
conv1d_68 (Conv1D)	(None, 11, 512)	393728
max_pooling1d_38 (MaxPooling1D)	(None, 5, 512)	0
conv1d_69 (Conv1D)	(None, 5, 35)	53795
max_pooling1d_39 (MaxPooling1D	(None, 2, 35)	0
lstm_5 (LSTM)	(None, 64)	25600
dense_27 (Dense)	(None, 64)	4160
dense_28 (Dense)	(None, 64)	4160
dense_29 (Dense)	(None, 1)	65
Total parameters: 612,612		
Trainable parameters: 612,612		
Non-trainable parameters: 0		

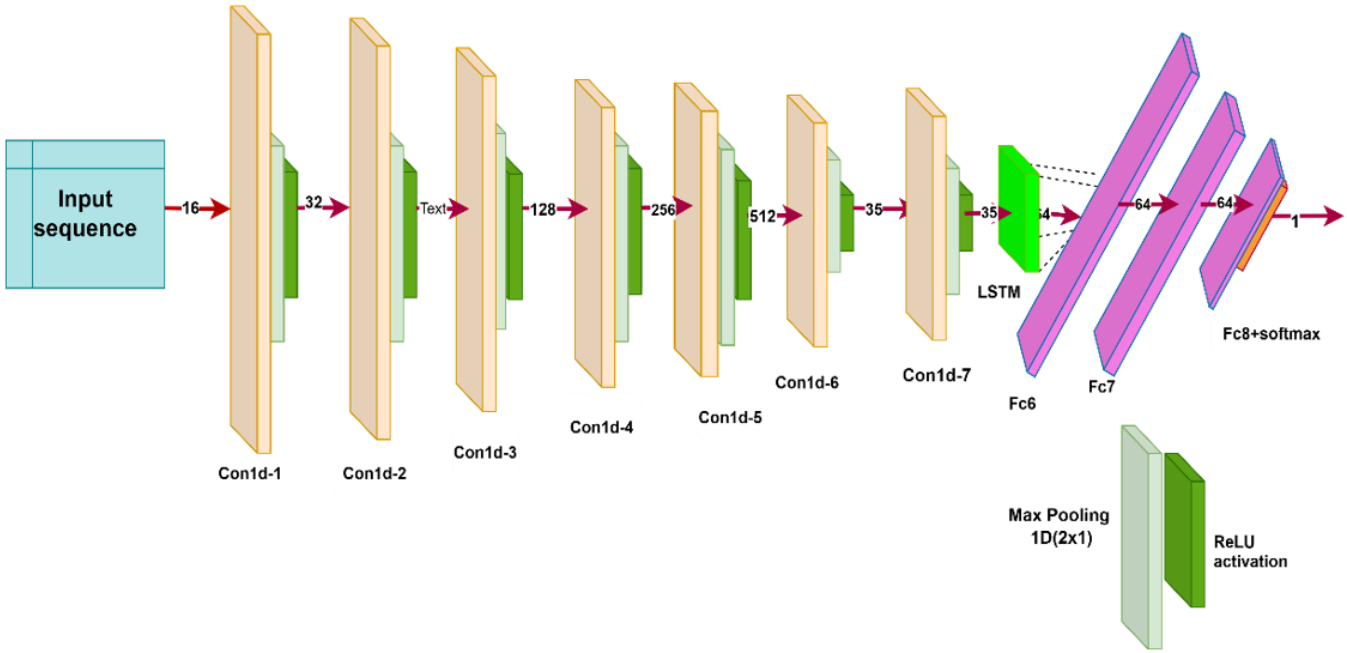

Figure 3. Block diagram of the suggested technique framework

6. EXPERIMENTS AND FINDINGS

The experiments and findings for each of the models put forward in the preceding part are summarized in this section. On two datasets, one of which included a dataset of Arabic spam and Ham tweets [41], which is globally measured, and the other on a dataset that was built and compiled from Twitter based on the Aho-Corasick algorithm with the scrape Python library for data scraping, which includes Several Arabic tweets crime editorials were used to detect crimes of Tweeter tweets to study the effect of features on model learning and the resulting accuracy.

6.1 Environmental settings

The research is conducted on an Intel Core i7 computer with Python 3.11.4, Google Collab, the Keras library for deep learning models, and Sklearn for machine learning techniques for comparisons.

6.2 Experiment 1 compares results with other ML and DL models

The proposed techniques are compared to various ML algorithms and deep learning on a 20% dataset and created data set. The technique assesses classifier performance using accuracy, precision, recall, and f1-measure variables. The best classifier is picked based on higher measurement values. These parameters are Precision (Eq. (3)), Recall (Eq. (4)), Accuracy (Eq. (5)), and F1-Score (Eq. (6)) [42].

$$\text{Precision} = \frac{\text{TP}}{TP + FP} \tag{3}$$

$$\text{Recall} = \frac{\text{TP}}{TP + TN} \tag{4}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{TP + TN + FP + FN} \tag{5}$$

$$\text{F1score} = \frac{2 \times \text{Precision} \times \text{Recall}}{Precision + Recall} \quad (6)$$

The study focuses on the efficiency and performance of classification methods using evaluation measurements. The dataset includes keywords for crimes, user names, tweets, and other features. The proposed method categorizes the dataset using ML classifiers and DL models to select the best classifier for better-assessed metrics. The results were compared to a dataset of (13240) tweets [41], and compared with a paper [12] model using the same data set. The outcomes of the DL and ML were as follows. The hybrid CNN-LSTM model achieved the highest accuracy of 99.84, and the results for RF using f1-measure is 98.76, an accuracy of 100. The results also show that the machine learning of the model outperforms most classification techniques in terms of accuracy and precision. Table 4 illustrates measurements of the ML and DL models' performance. The table shows that the proposed approach produced better outcomes than the proposed method in the paper [12]. While the paper [12] utilizing ML, and DL techniques (Fast Text+LSTM) and (GloVe+LSTM) yielded the following results for the Text+LSTM, the classifier obtained with no computation of overall accuracy, the total classifier precision, recall, and f1-measure are 99.1 92.4 2, and 95.1, respectively. The classifier obtains for the GloVe+ LSTM The total classifier precision, recall, and f1-measure are 98.88, 95.92, and 97, respectively, with no measurement of overall accuracy, but machine learning outperformed most classification algorithms in terms of precision.The results show high-performance classifiers, effectively processing text data through pre-processing steps like tokenization, stop word removal, derivation, and text normalization. These models provide accurate and valid classification based on pre-processed data. The second group (the constructed data set, which totaled 18493), and the findings are shown in Table 5: The proposed method for classifying Twitter Crime Detection performances.

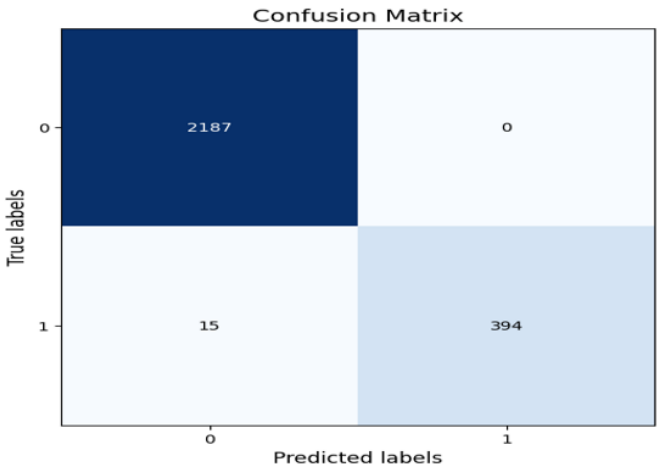

Figure 4. Confusion matrix for XGBoost model based on the dataset [41]

When compared to F1, the findings showed that the Hybrid CNN-LSTM model performed the best. Figure 4 depicts for XGBoost Model machine learning classifier.

When comparing the performance of ML and DL algorithms, the test results from the data set that it built are efficient and have high performance in all machine learning categories of crime and deep learning, As the results of the XGBoost model, the classifier achieves an accuracy of 100, while the overall classifier precision, recall, and f1-measure are 99, 99.63, and 99.36, respectively. Table 6 summarizes the Twitter Crime Detection classifiers with the highest accuracy, recall, and F1 measurement findings, all of which are higher than the paper [12] using Dataset [41]. XGBoost measurement scores are greater in terms of accuracy, precision, recall, and f1. The Hybrid CNN-LSTM classifier has an overall accuracy of 99.75. The total precision, recall, and f1-measure of the classifier are 98.76, 98.75, and 99.86, respectively. According to the results, the Hybrid CNN-LSTM has a higher accuracy of 99.75 than the F1-Score. Figure 5 depicts the Machine Learning Confusion Matrix.

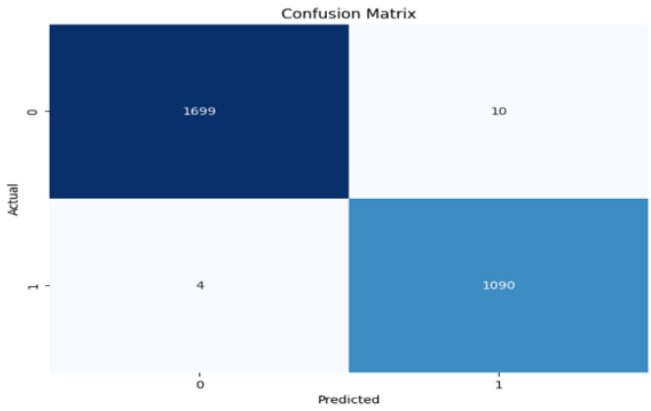

Figure 5. Confusion Matrix for XGBoost Model based on data set built

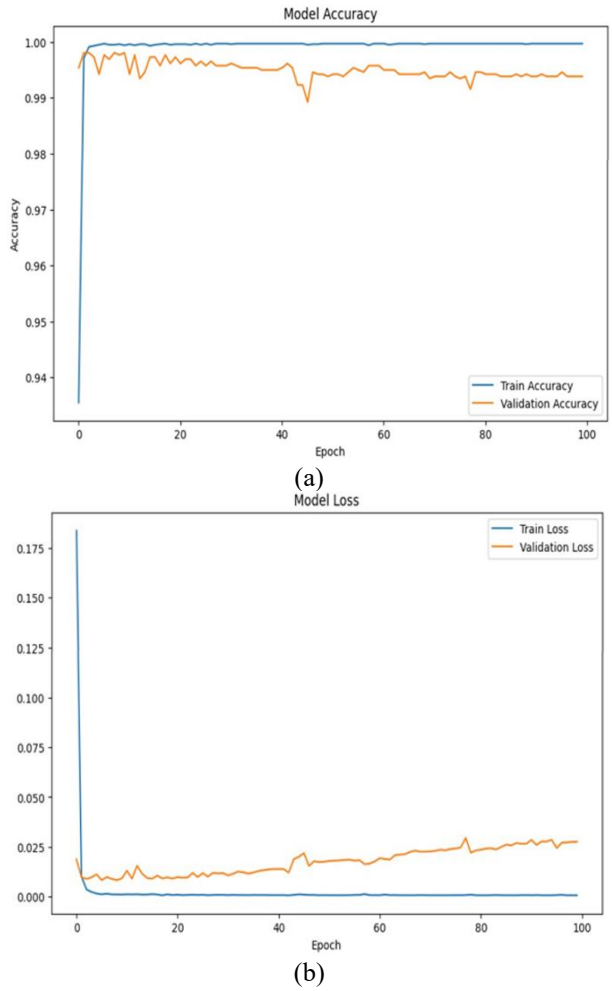

Figure 6. Performance of proposed model (a) accuracy curve (b) loss curve

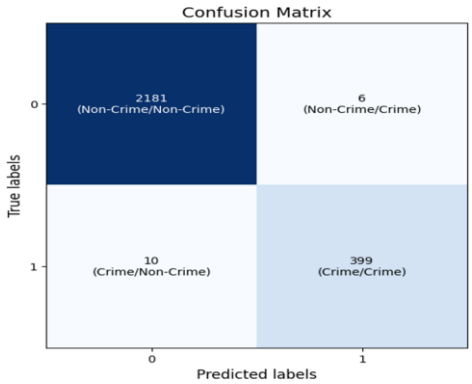

Figure 7. Confusion matrix of hybrid CNN-LSTM model of crimes tweets

Compare the proposed model's performance to additional measures such as Mean Average Precision (MAP) and Average Precision for each class, as given in Table 6: Deep Learning Performance Model Comparison structure.

The Hybrid CNN-LSTM model achieved a 99.84% accuracy rate in detecting tweet crime for 100 epochs and batch size = 64, as shown in Figure 6, and the loss and accuracy model and confusion matrix are illustrated in Figure 7.

6.3 Experiment 2 compares proposed algorithms with the literature

This section illustrates compare in Table 7 displays the best results of proposed algorithms compared to the literature.

Table 4. Measurements of the ML and DL models' performance

Dataset of Arabic Spam and Ham Tweets (13240) [41] Based on Proposed Method				
Model for the proposed Method	**Precision**	**Recall**	**F1 Score**	**Accuracy**
XGBoost+TF-IDF	100	96.33	98.13	99.4
	Model for DL			
Hybrid CNN-LSTM Model	**99.84**	99.84	**99.84**	**99.84**
The Result of the Paper [12] Based on the Dataset [41]				
Model for paper [12]	**Precision**	**Recall**	**F1 Score**	
Naïve Bayes Models with SMOTE	98.43	98.45	98.94	
SVM with SMOTE	99.8	99.19	**99.49**	
LR with SMOTE	99.4	99.48	**99.44**	
Neural Network without SMOTE	99.98	99.48	**99.73**	
Model for DL				
Fast Text+LSTM	99.1	92.42	**95.1**	
GloVe+LSTM	98.88	95.92	**97**	

Table 5. The proposed method for classifying Twitter Crime Detection performances

Dataset of Arabic Spam and Ham Tweets (13240) [41] Based on Proposed Method				
Model for ML	**Precision**	**Recall**	**F1-Score**	**Accuracy**
XGBoost	**100**	96.33	98.13	99.4
Model for DL				
Hybrid CNN-LSTM	**99.84**	99.84	**99.84**	**99.84**
A dataset of Arabic crime tweets (18493) built with the proposed method				
Model for ML	**Precision**	**Recall**	**F1-Score**	**Accuracy**
XGBoost	99	**99.63**	**99.36**	**100**
Model for DL				
CNN-LSTM	98.76	98.75	**99.86**	**99.75**

Table 6. Deep learning performance model comparison structure

Model	**Average Precision for Each Class**	**Mean Average Precision (MAP)**
Hybrid CNN-LSTM	99.82	99.82

Table 7. Comparison of proposed paper with related publications

Paper Authors	**Year of Publication**	**Data Source/ Language**	**The Best Performance of the Algorithm**	**F1 Score or Accuracy**
Kaddoura et al. [12] and dataset [41]	2023	Tw/ Arabic	NN (No SMOTE)	**99.73%**
Mohaouchane et al. [13]	2019	YT/ Arabic	CNN	87.8%
Abuzayed et al. [14]	2020	Tw/ Arabic	CNN-LSTM	73%
Muneer et al. [15]	2021	Tw/EN	IDCNN with BiLSTM	96.4%
Aref et al. [16]	2020	Tw/ Arabic	CNN, Fast text	52%
Aljarah et al. [17]	2020	Tw/ Arabic	RF	91.3%
Method Proposed	2023	Tw/ Arabic	Hybrid_CNN LSTM model	**99. 84%**

7. CONCLUSIONS

In conclusion, our study effectively used the Hybrid_CNN LSTM model and the XGBoost machine learning model to create a reliable Twitter Intelligent Criminal Detection system. The suggested intelligent dictionary model employs the Aho-Corasick algorithm to identify criminal tweets in Arabic. Because of the increased accuracy, this technology is useful for improving public safety in the digital age. Further investigation and cooperation with law enforcement are needed to perfect and use this cutting-edge criminal detection technology. The model's performance can be evaluated using various metrics, such as accuracy, precision, recall, F1 score, and APM, to assess its effectiveness in detecting crime-related tweets in Arabic. The outcomes demonstrated deep learning model efficacy and superiority over other methods. The results for the Hybrid CNN LSTM model achieved an accuracy of 99.75% and an F1 macro score of 99.84%. The conventional ML models outperformed all others with an F1 macro score of 99.36% and an accuracy level of 100% when trained using the XGBoost approach. This proves that DL approaches find offensive tweets in the given dataset better. Hybrid CNN-LSTM also achieved a mean average precision (MAP) of 99.82 when tested against other metrics, demonstrating the efficacy of the deep learning model in the suggested approach. Research in the future can concentrate on enhancing the model's capacity to react to new trends and changing patterns of criminal conduct on Twitter in Arabic, adding domain-specific vocabularies, and broadening the dictionary's coverage. To further improve the model's accuracy and robustness, it is recommended to enhance preprocessing techniques and explore novel deep learning architectures continuously. The suggested approach proves that an intelligent dictionary can find Arabic tweets about crimes. An effective approach to identifying and addressing criminal content is provided by combining machine learning, deep learning, and the Aho-Corasick algorithm. This contributes to a safer online environment.

ACKNOWLEDGMENT

The authors would like to thank the Informatics Institute for Postgraduate Studies, Iraqi Commission for Computers and Informatics (https://iips.edu.iq/), Baghdad-Iraq for its support in the present work.

REFERENCES

[1] Simović, M., Kuprešanin, J. (2020). Criminal offenses with elements of violence-psychology of crime and abuse of power. Knowledge - International Journal, 42(5): 933-938. https://ikm.mk/ojs/index.php/kij/article/view/648

[2] Farrall, S., Gray, E., Mike Jones, P. (2020). Politics, social and economic change, and crime: Exploring the impact of contextual effects on offending trajectories. Politics & Society, 48(3): 357-388. https://doi.org/10.1177/0032329220942395

[3] Mccord, M., Chuah, M. (2011). Spam detection on twitter using traditional classifiers. In Autonomic and Trusted Computing: 8th International Conference, ATC 2011, Banff, Canada, pp. 175-186. https://doi.org/10.1007/978-3-642-23496-5_13

[4] Alkhalil, Z., Hewage, C., Nawaf, L., Khan, I. (2021). Phishing attacks: A recent comprehensive study and a new anatomy. Frontiers in Computer Science, 3: 563060. https://doi.org/10.3389/fcomp.2021.563060

[5] Greer, C. (2013). Crime and media: understanding the connections. Criminology, 3: 143-164.

[6] Adam, I., Fazekas, M. (2021). Are emerging technologies helping win the fight against corruption? A review of the state of evidence. Information Economics and Policy, 57: 100950. https://doi.org/10.1016/j.infoecopol.2021.100950

[7] Yang, D., Heaney, T., Tonon, A., Wang, L., Cudré-Mauroux, P. (2018). CrimeTelescope: Crime hotspot prediction based on urban and social media data fusion. World Wide Web, 21: 1323-1347. https://doi.org/10.1007/s11280-017-0515-4

[8] Ristea, A., Leitner, M. (2020). Urban crime mapping and analysis using GIS. ISPRS International Journal of Geo-Information, 9(9): 511. https://doi.org/10.3390/ijgi9090511

[9] Zhong, Z., Bao, Y., Shen, S., Zhou, E. (2020). Predict New app quality by using machine learning. In Journal of Physics: Conference Series, 1693(1): 012111. https://doi.org/10.1088/1742-6596/1693/1/012111

[10] Guellil, I., Adeel, A., Azouaou, F., Benali, F., Hachani, A.E., Dashtipour, K., Hussain , A. (2021). A semi-supervised approach for sentiment analysis of arab (ic+ izi) messages: Application to the algerian dialect. SN Computer Science, 2: 1-18. https://doi.org/10.1007/s42979-021-00510-1

[11] Sharma, N., Sharma, R., Jindal, N. (2021). Machine learning and deep learning applications-a vision. Global Transitions Proceedings, 2(1): 24-28. https://doi.org/10.1016/j.gltp.2021.01.004

[12] Kaddoura, S., Alex, S.A., Itani, M., Henno, S., AlNashash, A., Hemanth, D.J. (2023). Arabic spam tweets classification using deep learning. Neural Computing and Applications, 35: 17233-17246. https://doi.org/10.1007/s00521-023-08614-w

[13] Mohaouchane, H., Mourhir, A., Nikolov, N.S. (2019). Detecting offensive language on arabic social media using deep learning. In 2019 sixth international conference on social networks analysis, management and security (SNAMS), pp. 466-471. https://doi.org/10.1109/snams.2019.8931839

[14] Abuzayed, A., Elsayed, T. (2020). Quick and simple approach for detecting hate speech in Arabic tweets. In Proceedings of the 4th workshop on open-source Arabic Corpora and processing tools, with a shared task on offensive language detection, pp. 109-114. https://aclanthology.org/2020.osact-1.18

[15] Muneer, A., Fati, S.M. (2020). A comparative analysis of machine learning techniques for cyberbullying detection on twitter. Future Internet, 12(11): 187. https://doi.org/10.3390/fi12110187

[16] Aref, A., Al Mahmoud, R.H., Taha, K., Al-Sharif, M. (2020). Hate speech detection of Arabic shorttext. In CS IT Conference Proceeding, 10: 81-94. https://doi.org/10.5121/csit.2020.100507

[17] Aljarah, I., Habib, M., Hijazi, N., Faris, H., Qaddoura, R., Hammo, B., Alfawareh, M. (2021). Intelligent detection of hate speech in Arabic social network: A machine learning approach. Journal of Information Science, 47(4):

483-501. https://doi.org/10.1177/0165551520917651
[18] Twitter Developer. (2021). https://developer.twitter.com/en/docs/twitter-api/v1/rate-limits (accessed on 1 November.
[19] Ourlis, L., Bellala, D. (2019). SIMD implementation of the aho-corasick algorithm using intel AVX2. Scalable Computing: Practice and Experience, 20(3): 563-576. https://doi.org/10.12694/scpe.v20i3.1572
[20] Sheikh, H., Prins, C., Schrijvers, E. (2023). Artificial intelligence: Definition and background. In Mission AI: The New System Technology, pp. 15-41. https://doi.org/10.1007/978-3-031-21448-6_2
[21] Chandra, Y., Jana, A. (2020). Sentiment analysis using machine learning and deep learning. In 2020 7th international conference on computing for sustainable global development (INDIACom), pp. 1-4. https://doi.org/10.23919/INDIACom49435.2020.9083703
[22] Shubham, D., Mithil, P., Shobharani, M., Sumathy, S. (2017). Aspect level sentiment analysis using machine learning. In IOP Conference Series: Materials Science and Engineering, 263(4): 042009. https://doi.org/10.1088/1757-899X/263/4/042009
[23] Hua, T., Chen, F., Zhao, L., Lu, C.T., Ramakrishnan, N. (2013). STED: Semi-supervised targeted-interest event detectionin in twitter. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1466-1469. https://doi.org/10.1145/2487575.2487712
[24] Pandarachalil, R., Sendhilkumar, S., Mahalakshmi, G.S. (2015). Twitter sentiment analysis for large-scale data: an unsupervised approach. Cognitive Computation, 7(2): 254-262. https://doi.org/10.1007/s12559-014-9310-z
[25] Cai, J., Hao, J., Yang, H., Zhao, X., Yang, Y. (2023). A review on semi-supervised clustering. Information Sciences, 632: 164-200. https://doi.org/10.1016/j.ins.2023.02.088
[26] Shinde, P.P., Shah, S. (2018). A review of machine learning and deep learning applications. In 2018 Fourth international conference on computing communication control and automation (ICCUBEA), pp. 1-6. https://doi.org/10.1109/ICCUBEA.2018.8697857
[27] Salim, K., Hebri, R.S.A., Besma, S. (2022). Classification predictive maintenance using XGboost with genetic algorithm. Revue d'Intelligence Artificielle, 36(6): 833-845. https://doi.org/10.18280/ria.360603.
[28] Alsaedi, E. (2022). A comparative study of combining deep learning and homomorphic encryption techniques. Al-Qadisiyah Journal of Pure Science, 27(1): 17-33. https://doi.org/10.4018/IJCAC.309936
[29] Liu, F., Xue, S., Wu, J., Zhou, C., Hu, W., Paris, C., Yu, P.S. (2020). Deep learning for community detection: progress, challenges and opportunities. arXiv preprint arXiv:2005.08225. https://doi.org/10.24963/ijcai.2020/693
[30] Ravì, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., Yang, G.Z. (2016). Deep learning for health informatics. IEEE Journal of Biomedical and Health Informatics, 21(1): 4-21. https://doi.org/10.1109/JBHI.2016.2636665
[31] Mosavi, A., Ardabili, S., Varkonyi-Koczy, A.R. (2019). List of deep learning models. In International conference on global research and education, pp. 202-214. https://doi.org/10.1007/978-3-030-36841-8_20
[32] Yao, Y., Huang, Z. (2016). Bi-directional LSTM recurrent neural network for Chinese word segmentation. In Neural Information Processing: 23rd International Conference, ICONIP 2016, Kyoto, Japan, October 16-21, 2016, Proceedings, Part IV 23, pp. 345-353. https://doi.org/10.48550/arXiv.1602.04874
[33] Kamath, U., Liu, J., Whitaker, J. (2019). Deep learning for NLP and speech recognition (Vol. 84). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-030-14596-5
[34] Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882. https://doi.org/10.3115/v1/D14-1181
[35] Oleiwi, B.K., Abood, L.H., Farhan, A.K. (2022). Integrated different fingerprint identification and classification systems based deep learning. In 2022 International Conference on Computer Science and Software Engineering (CSASE), pp. 188-193. https://doi:10.1109/CSASE51777.2022.9759632
[36] Shehu, Y.I., Ruiz-Garcia, A., Palade, V., James, A. (2018). Sokoto coventry fingerprint dataset. arXiv preprint arXiv:1807.10609. https://arxiv.org/abs/1807.10609
[37] Alhussan, A.A., Farhan, A.K., Abdelhamid, A.A., El-Kenawy, E.S.M., Ibrahim, A., Khafaga, D.S. (2023). Optimized ensemble model for wind power forecasting using hybrid whale and dipper-throated optimization algorithms. Frontiers in Energy Research, 11: 1174910. https://doi.org/10.3389/fenrg.2023.1174910
[38] Wang, Y., Bao, D., Qin, S.J. (2023). A novel bidirectional DiPLS based LSTM algorithm and its application in industrial process time series prediction. Chemometrics and Intelligent Laboratory Systems, 104878. https://doi.org/10.1016/j.chemolab.2023.104878
[39] 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
[40] 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
[41] Kaddoura, S., Henno, S. (2023). Dataset of Arabic Spam and Ham Tweets. Mendeley Data, V1, https://doi.org/10.17632/86x733xkb8.2
[42] 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