Global Journal of Engineering and Technology Advances
eISSN: 2582-5003
Cross Ref DOI: 10.30574/gjeta
Journal homepage: https://gjeta.com/

(REVIEW ARTICLE)

Check for updates

Analysis of social networks and filtering of Arabic crime tweets based on an intelligent dictionary using a genetic algorithm

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

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

Global Journal of Engineering and Technology Advances, 2024, 18(02), 177–191

Publication history: Received on 11 January 2024; revised on 24 February 2024; accepted on 26 February 2024

Article DOI: https://doi.org/10.30574/gjeta.2024.18.2.0033

Abstract

Preserving a robust online community poses a significant difficulty due to the unrestricted flexibility members have in expressing themselves and behaving. This issue can be remedied through the implementation of user behavior monitoring and analysis, followed by the implementation of suitable actions. The objective of this research is to develop an intelligent dictionary using the genetic algorithm to identify and categorize Twitter posts related to criminal activities by collecting data from tweets. Once the data is preserved, graph analysis techniques are employed to evaluate interactions between users. Next, the user behavior is analyzed using metadata analysis, whereby the chronology associated with each user profile is obtained. Furthermore, the study analyzes the behavioral patterns of users over time. Afterward, a method based on rules is used to create a structure for aspect-based analysis of sentiment. This method assesses the subjectivity of the input text, distinguishing between factual information and personal opinions. Additionally, transformer-based sentiment analysis determines whether the tweet evokes positive or negative sentiment. Furthermore, the task involves constructing a model that can accurately categorize a tweet based on its relevance to criminal activity. Ultimately, the intelligent dictionary is utilized to identify and isolate anomalous behavior by selecting provocative profiles. A Twitter profile exhibiting significant similarity with criminal cases presents a perilous risk to society.

Keywords: Cybercrime; Social network analysis; Twitter analysis; Natural Language Processing (NLP); Genetic Algorithm.

1 Introduction

Security is crucial for human life, and with the rise of internet users and mobile data technology, there is a growing volume of information-related crimes. Unstructured data, such as "free text," has led to the need for methods to manage it. Social media platforms like Facebook, Twitter, and Snapchat have become popular for private messages, sharing photos, and sharing news. Twitter, with around 500 million users, has become an important means of communicating criminal activity news due to its small, readable tweets [1]. Social media monitoring can enhance crime reporting by facilitating criminal activity, similar to other emerging technologies and communication channels [2]. Twitter is different from other social media sites because it lets people share news, thoughts, and ideas in just 280 characters. On other sites, links between texts are not shared [3]. Twitter's tracking and tweeting capabilities enable researchers to identify cyber risks in real time, utilizing artificial intelligence for security intelligence and criminal behavior analysis[4]. Crime has varying effects on society, and social media serves as a medium for worldwide communication. Nevertheless, perpetrators might selectively victimize someone based on their race, physical appearance, or religious convictions [5]. Crime engenders adverse consequences for both victims and non-victims, manifesting as

* Corresponding author: Zainab Khyioon Abdalrdha

Copyright © 2024 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution Liscense 4.0.

diminished security, reduced work productivity, financial losses, property damage, and medical ailments. Governments allocate funds towards the provision of protection, courts, therapeutic courses, and social workers. Crime detection can contribute to the development of a more robust society and enhanced well-being, by mitigating distress and fostering economic prosperity [6]. This study employs a range of centrality measures, such as degree, closeness, betweenness, eigenvector, page rank, and weight, to examine tweet owners and tweets to create a directed social network graph. Since connections can happen from only one side of a directed network, it is appropriate for describing Twitter data in this way. This allows for the analysis of user relationships, as directed graphs do not require symmetric ties between vertices. In addition, identify information sources and influencers in social networks based on 18,493 tweets collected and built into the paper[7]. Additionally, the study offers a visualization of every community and social network. For Arabic crime tweets, word cloud analysis of users' interests and subjects is used. To the best of our knowledge, this is the first study to use a social network analytic approach to investigate Arabic crime tweets. This study's primary contributions:

- A social network visualization for crime tweets.
- Identification of information sources used in Arabic criminal tweets shared on Twitter.
- Identifying the most influential individuals in Arabic criminal tweets.
- Use a genetic algorithm dictionary to identify and filter abnormal Arabic crime behavior.

By identifying Twitter's primary sources of information and influencers and filtering anomalous behavior in Arabic crimes, security officials can assess tweeters' knowledge. In addition to utilizing specified sources of information and influencers, the government may also distribute guidelines and instructions to mitigate and eliminate risks. The organization of this paper is as follows: Section 2 provides an overview of the related works. Section 3 outlines the proposed method. Section 4 presents the experiments and findings. Finally, Section 5 presents the conclusions and future work.

2 Related work

Research has used data and artificial intelligence (AI) to discover, improve, and predict events, including anomaly detection [8] Information recovery, analysis, and profile generation systems [9]. And the like: exploring crime-related tweets using social network visualization and extracting user sentiment from Twitter for crime detection is the main area of research at the moment. In the fields of travel advertising, event detection, criminal detection, and human behavior, this study aims to discover action-driven patterns. What follows is a summary of some current research on this topic. The authors in this paper[10], propose a crime prediction model using historical data and spatiotemporal data by crime type. The model uses a vector of n values, with a neural network algorithm providing an 81% accuracy, making it faster and more suitable for the data format. The study [11], employed Twitter data from seven specific areas in India to examine crime patterns and geographical dispersion. The analysis of real-time crime data was conducted using machine learning algorithms and natural language processing techniques. The tracking of criminal activity was conducted using sentiment analysis and Brown clustering. The dataset, comprising 11,073 cybersecurity tweets, underwent evaluation using four classifiers. The accuracy rate of the Random Forest model was 98.1%, which was the highest among all models. On the other hand, the ZeroR model had the lowest accuracy rate of 61.5%. This study[12], used Twitter data to categorize users using a feature-based approach, combining graph and metadata analysis. It determined the importance of nodes, identifying Influencers and Fakes. The authors of this work[13], analyzed the data they obtained to establish the crime rate in various locations. They utilized the K-nearest neighbor algorithm and employed various data-gathering approaches to analyze the crime trend in Bangladesh from 2017 to 2019. The K-nearest neighbor algorithm yields the highest precision for the crime rate forecast system, achieving an accuracy of 76.9298%. The study in this paper [14], used text-mining techniques to categorize tweets about criminal incidents necessitating police action. The system employs classifiers such as Naive Bayesian, Random Forest, J48, and ZeroR. Data was gathered from seven Indian towns spanning the years 2014 to 2016. The method yielded an F1-score of 94.92, a classification accuracy of 94.91%, a precision of 94.94%, and a recall accuracy of 93.91%. These results demonstrate that the technique is an effective tool for detecting illegal activities. The study [15], utilized text-mining methods to classify 369 tweets as crime and non-crime. Machine learning algorithms were used to identify Twitter accounts that necessitate police involvement. The labeled data was used to train Naïve Bayesian and other models. The Random Forest model attained the highest accuracy, achieving a 98.1% accuracy rate. The paper [16] used dynamic time-warping to analyze profiles of users, focusing on behavioral habits, audience, and shared content, and determining similarity using the TF-IDF and cosine similarity metrics. This study [17] analyzes tweet data to create a Twitter monitoring tool. Graph analysis examines user behavior following data maintenance. The chronological data of each profile and temporal patterns of behavior are used in metadata analysis to evaluate user activity. The component that identifies and filters aberrant behavior selects intriguing profiles for further analysis using nine filters. The contextual analysis component uses a binary text classification model with SVM and TF-IDF. This model detects crimes in tweets with 88.89% accuracy. We use aspect-based sentiment analysis with 80% accuracy using Distil BERT and FFNN (Feed-Forward Neural

Network) to evaluate tweets about unlawful acts to reduce the harm caused by positive attitudes about crime. This platform advises law enforcement on fighting hate speech and terrorism.

3 The proposed mothed

The suggested approach sought to identify profiles with a high degree of criminal activity. As shown in figure 1. The main target for investigating crime and analysis is the detection of abnormal behavior in Twitter tweets. It allows interaction in two directions, making it simple and quick for any user to engage with another. Users can influence the opinions of a group of people by publishing important information and using various content-sharing techniques. In addition, This section describes the remaining procedures of the main framework for the suggested approach.

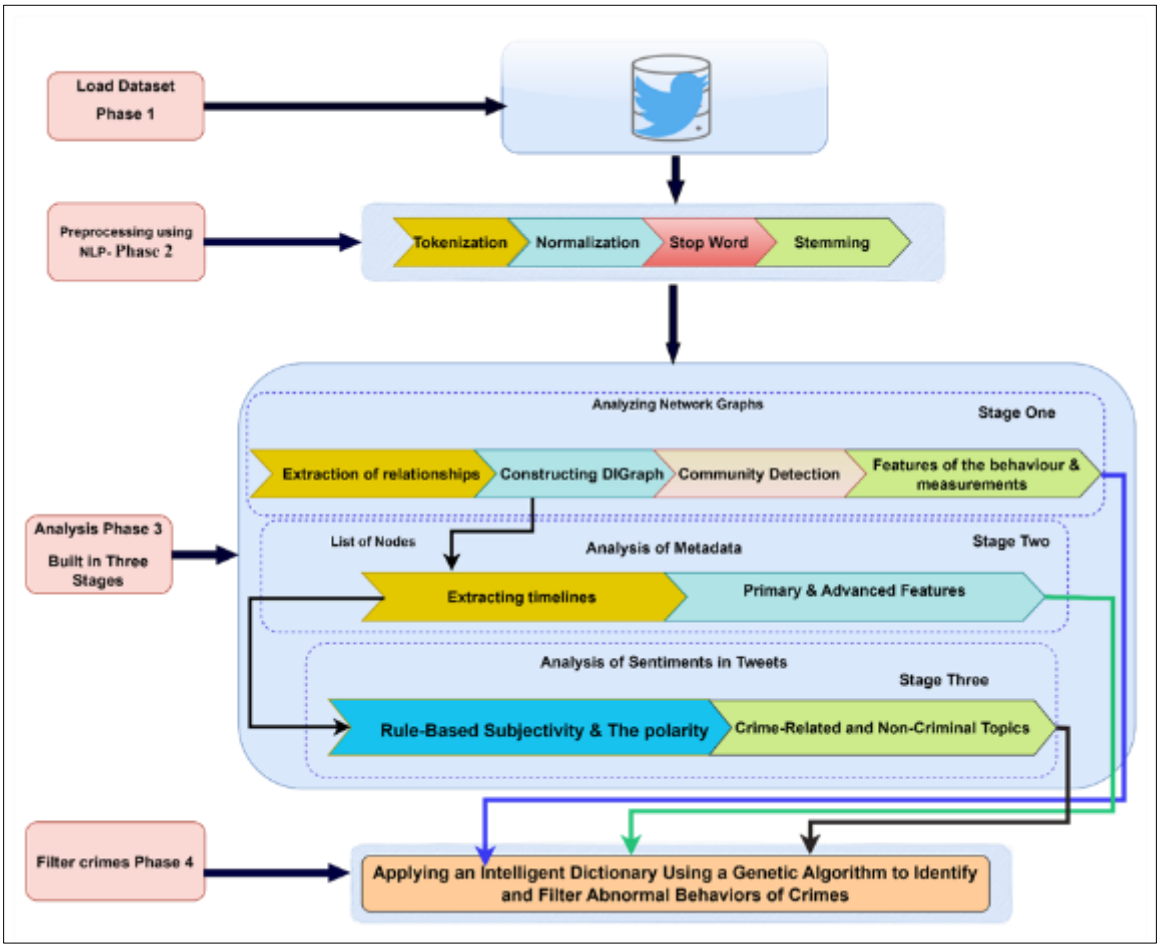

Figure 1 The methodology of the behaviors of Arabic crime Tweets

3.1 Database of Arabic Crime Tweets

Although the Twitter development team provides an official Twitter API [18], Python offers several library utilities for data extraction, such as Tweepy and Scrape. The present study utilizes a dataset constructed utilizing an intelligent dictionary [7], to extract and analyze crime-related information. The Gephi software was employed to analyze the dataset (https://gephi.org). Free and open-source software intended for investigating and visualizing networks and graphs is known as Gephi. Its primary purpose in our study was to investigate and implement social network analysis for the detection of abnormal behavior in Twitter tweets. This paper analyzed a total of 18,493 data sets.

3.2 Preparing and Cleaning Tweets

The study uses a dataset of 18,493 tweets for crime tweet detection, preprocessing them using normalization, tokenization, stop-word removal, and stemming techniques. Normalization reduces noise, tokenization converts characters into tokens for linguistic analysis, stop-word removal removes unnecessary words, and stemming removes prefixes and suffixes from inflected words, ensuring uniform classification across all datasets[19]. Then to achieve precise visualization of social networks, it is necessary to do data preprocessing. As a result, the tweets were filtered and any unnecessary ones were removed from the study once the data was loaded into the Gephi software.

3.3 Analysis phases

In this section, we will elucidate the crucial steps of social media analysis specifically for Twitter. This section is comprised of three distinct stages: the network analysis stage, the metadata stage, and the sentiment analysis stage. These stages are conducted using Twitter's tweets, and we will provide a detailed description of each below.

3.4 Analyzing Network Graphs

Initially, the network analysis component generated a node list, which was a compilation of the distinct users, by extracting the interactions between individuals. Following the extraction of data, the users' intercommunication was identified, and a corresponding graph network was constructed. This paper will examine a subset of user interconnections. The nodes are composed of user identities and tweets, while likes, reposts, replies, quotes, followers, and interactions represent the edges. The graph centrality measures show that profiles with more frequent appearances are more significant, as they have a greater impact on the network. As shown in Figure 2, the most common usernames.

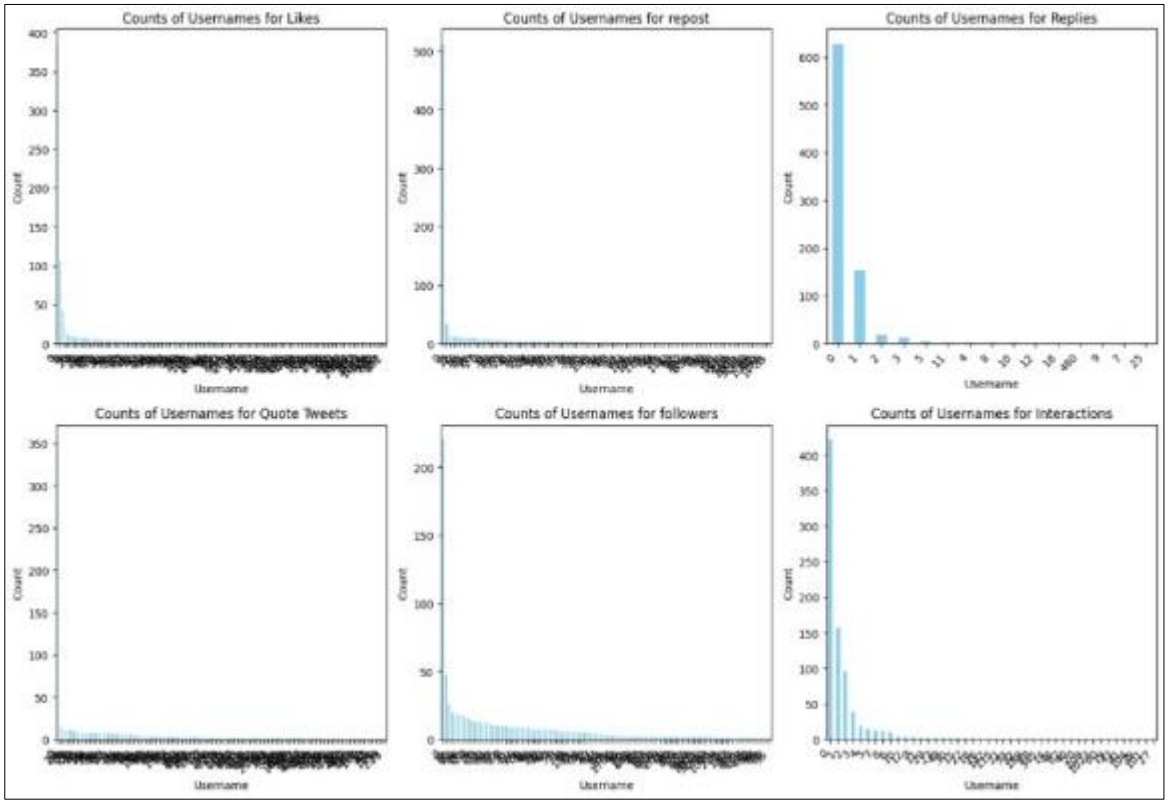

Figure 2 The most common usernames**.**

After assessing the network's features, the community discovery method is executed. A comprehensive understanding of communities and nodes is established. In this paper, two strategies are employed to detect communities. Community detection algorithms can discern user clusters based on their quoting, mentioning, and retweeting behaviors. Community discovery methodologies are divided into four categories: graph partitioning, spectral clustering, modularity, and label propagation. The initial three options utilize recursive partitioning or sophisticated network unions. Complex networks exhibit hierarchical community structures. The Girvan-Newman algorithm detects communities in graphs, whether they are directed or undirected. This methodology utilizes the technique of partitioning graph clustering. Although it is commonly utilized, its scalability and computational complexity for weighted graphs are O (m^3) for weighted graphs and O((m^3) + (m^3)log m) for unweighted graphs[20]. The Network X library was used to create Algorithm 1. [21].

During the social network analysis phase, a directed social network graph was constructed by analyzing and visualizing the tweet proprietors. A limited number of algorithms were implemented to enhance the network's visibility and derive valuable insights from the visualization of the connections between tweet proprietors. A fundamental attribute of social

network analysis is the ability to identify influential and prominent entities within the network. Centrality measures are frequently employed as indices that are constructed using network data. Typically, they signify the significance of a node, encompassing attributes such as prestige, visibility, structural strength, or status.

Algorithm 1 Girvan-Newman Algorithm

Input: Graph network with directed edges
Output: Nodes and community numbers matrix
1. Network Edge Betweenness Calculation
- Initial calculation of each edge's betweenness.
2. Elimination of Edge(s) with Largest Betweenness Centrality
3. After they are removed, the betweenness of every impacted edge is calibrated again.
4. The process of repeating steps 2 and 3 continues until there are no remaining edges.

Popular measures of central tendency include degree, proximity, betweenness, eigenvector, and page rank A concise explanation of these measures is provided below[17].

- **Eccentricity**: The greatest minimum distance between one node and all others. A node with reduced eccentricity possesses a higher degree of impact over others.
- **Clustering Coefficient Centrality:** A group of nodes in a network that has a strong tendency to cluster together according to their degree. $cc = \frac{m}{t}$, the symbol (cc) indicates clustering coefficient centrality, while the symbol m indicates the total number of connections between associates of a particular vertex, and the symbol (t) indicates the number of possible connections between all vertex members.
- **Closeness Centrality:** The average distance between vertices in a network indicates how close a node is to others. $Ccl = \frac{1}{\sum_{v \neq u} d(u,v)}$, the symbol *(Ccl)* indicates Closeness centrality, while the symbol d(u, v) indicates The geodesic length refers to the shortest distance between vertices u and v along the edges..
- **Betweenness Centrality:** Illustrates the impact of nodes. A higher betweenness centrality value reflects the significance of a node in the shortest paths of a network. Deleting this node would lead to the loss of multiple connections. Cb=∑_(n≠v≠l)▒〖d (δ(w))/δnl〗 , The symbol (Cb) stands for Betweenness centrality, δnl for Number of shortest paths between vertices n and l, δnl (w) for shortest paths through vertex w, (n) for Origin, and (l) for End.
- **Harmonic Closeness:** This metric exhibits analogies to proximity centrality; nevertheless, it applies to unlinked networks. By substituting the harmonic mean for the average distance, it becomes possible to deduce harmonic proximity, which defines infinity as separating two unconnected nodes.
- **Degree:** The sum of a node's in-degree and out-degree values indicates the total number of edges connected to it, irrespective of the direction of the edges.
- **In-Degree:** This metric accurately measures the number of vertices that a specific node on the web traverses.
- **Out-Degree:** This quantity represents the exact number of vertices originating from a specific node in the web..
- **Eigenvector centrality:** Eigenvector centrality calculates a node's significance based on its neighbors. A matrix operation on the adjacency matrix determines the primary eigenvector. PageRank uses eigenvector centrality to rank websites.

3.5 Analysis of Metadata

The component was responsible for analyzing metadata. This involves extracting timelines, primary features, and advanced features. The dataset comprises Arabic tweets with X features, as illustrated in Table 1. This table displays the most significant features retrieved from the tweets.

Table 1 Feature extraction from tweets[7].

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.

A tweet can be of various types such as 'Likes', 'reposts',' replies',' quote tweets',' followers', and 'interactions'. Fig. 3 illustrates the number range of the distribution of users of the tweets in a dataset.

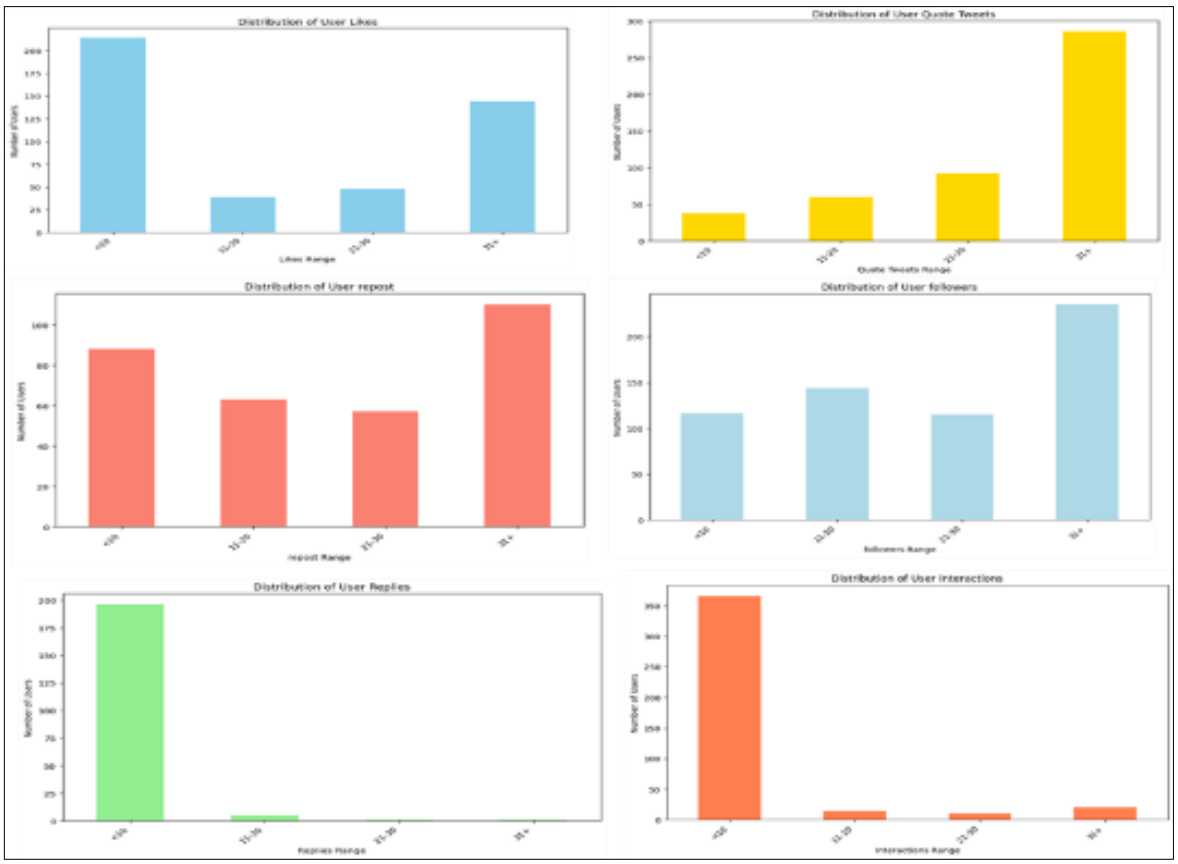

Figure 3 Number range of Distribution of User

The tweet object contains the generic information regarding a particular tweet, which is accessible in CSV format. This entity encompasses data about the tweet itself, including its text, creation date, number of favorites, retweets, and more. Additionally, It also includes the tweeter's followers, users followed, account creation date, total tweets, lists, and more. Following Twitter object data extraction [22]. The process involves extracting timelines and converting user attributes into time-series values, indicating consistency, activity level, and engagement. The data is grouped by day of publication to create a daily tweet series. This helps detect behavioral filtering, identifying influencers based on high interactions or outliers. Further analysis of profiles' content provides more information, enabling more targeted marketing strategies. This stage generates advanced features, as seen in Table 2.

Table 2 Advanced characteristics extracted from primary features relate to the profile's timeline**.**

Advanced Features Overview	Definition
Tweets of Original	The average amount of unique, non-quote, tweets posted.
Retweets	The average amount of retweets shared
Statuses	Average amount of posted statuses (including both original tweets and retweets)
Replies	The average amount of replies
Favorites	The average amount of likes received
Mentioned people	The typical number of persons whose names appear in the timeline postings
Hashtags	An average of how many hashtags were used
URL	The average amount of tweets that include a URL

3.6 Analysis of Sentiments in Tweets

This section describes sentiment analysis, which entails examining a text to discern its emotional content. There are numerous approaches to sentiment analysis [23], consisting of three primary approaches: methods based on rules, features, and embeddings[24- 27]. The process of analyzing sentiments in tweets involves the utilization of rule-based models to assess the subjectivity of an input text, distinguishing between factual statements and opinions. Sentiment analysis categorizes whether a tweet elicits positive, negative, or neutral emotions. The Text Blob method, implemented through a library called Text Blob, is utilized to quantify the subjectivity of the tweets. The tool enables part-of-speech tagging, sentiment analysis, categorization, word extraction, and others.. Text Blob's emotion is determined by the polarity and subjectivity of the input text. The concept of polarity refers to having opposite or contrasting qualities or characteristics. Sentiments are categorized into positive, negative, or neutral using a ternary conditional expression: a polarity larger than 0 indicates positivity, a polarity less than 0 indicates negativity, and a polarity equal to 0 indicates neutrality. Users view the subjectivity score text as objective or subjective, with 0.0 representing perfect objectivity and 1.0 representing full subjectivity. However, the subjectivity becomes apparent when determining whether the input text aligns more closely with an opinion or a fact. Subjectivity is categorized as either opinion or reality, depending on a threshold value of 0.5. If the subjectivity is greater than 0.5, it indicates the presence of an opinion. Conversely, if the subjectivity is less than or equal to 0.5, it is considered a fact. Additionally, Text Blob ignores unknown terms, analyzes words and phrases for polarity, and averages the scores.

3.7 Applying an Intelligent Dictionary Using a Genetic Algorithm to Identify and Filter Arabic Crime Behaviors of Crimes

The next stage is to use a filtering mechanism to find and recognize aberrant behavior in crimes. In this step, a genetic algorithm is used to generate a dictionary of users whose behaviors are particularly interesting. Furthermore, the extensive search area is limited by taking into account all of the data gathered from the graph network and metadata analyses. The dataset contains 18,493 tweets and was gathered using the Aho-Corasick algorithm approach. This technique effectively creates a vocabulary for searching a big text corpus. In this work, a total of 15,485 tweets were selected for analysis. The search area was subsequently narrowed to 3,228 profiles,, resulting in a 17.45%reduction. A Twitter profile that is heavily associated with criminal concerns poses a substantial threat. An intelligent dictionary is created to remove uninteresting profiles to reduce component input size in this study. Intelligent dictionaries are created by studying interactive graph network data and metadata-derived user behavioral patterns. The most important categories used to build the smart dictionary based on the genetic algorithm are explained below:

- **Old spreader**: An "old spreader." is an account that tweets a lot in one day, unlike its regular activity.
- **Influencer (I):** This sort of influencer has a low frequency of publishing, but their tweets have a substantial effect, characterized by the highly significant centrality of both in-degree and betweenness.
- **Spreader (RT):** This is a term used to describe a profile that, on a recent particular day, retweeted a large number of tweets, but rarely posted original content. The past behavior of the account does not show a consistent pattern**.**
- **Influencer (II):** This refers to an influencer who has a large number of followers, obtains a significant number of favorites, routinely retweets, and possesses high centrality values.

- **Constant Spreader**: A profile characterized by frequent mentions of multiple individuals and a significant number of followed profiles.
- New profiles with significant engagement: These are recently created profiles that consistently publish a high volume of content, resembling the activity level of a bot.
- **Influencers (III):** This refers to a group of individuals who have a significant number of favorites and retweets on their social media posts. They are also known for consistently publishing content and have high centrality metrics.

The component for detecting and filtering abnormal behavior finds individuals displaying suspect conduct and refers them to the next component for further examination. This entails scrutinizing their content and delving deeper into their posts.

3.7.1 A Genetic Algorithm[28]:

GA has been employed as an intelligent dictionary for filtering Arabic tweets. Behavioral categories are determined by applying a sophisticated lexicon to the values acquired in the preceding stages. The objective is to minimize the data to accelerate the procedure. Heuristic algorithms like the Genetic Algorithm (GA) work based on Darwin's idea of evolution. Figure 4 shows how GA uses natural selection and crossover to get the most optimal solution. Genetic algorithms randomly create populations. Every individual has a gene that solves a certain problem. Chromosomes encode these genes. To address the issue, a specialized objective function is formulated. Genetic Algorithm (GA) consists of three primary operators, namely: (i) Selection, (ii) Crossover, and (iii) Mutation. The process of selection involves the deliberate choice of individuals from the current generation, who will then be employed to contribute to the composition of the subsequent generation. The individuals with the greatest fitness values are selected at this point. It takes on the function of recognizing the parent's inside chromosomes, that is, the two best chromosomes that help create the better generation. Until the intended answer is found, this method is repeated. The steps of the genetic algorithm (GA) for building an intelligent dictionary, as shown in Algorithm 2.

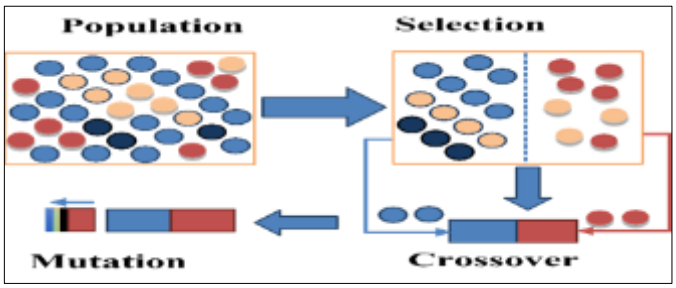

Figure 4 The Process of Implementing GA [29].

Algorithm 2 Build an Intelligent Dictionary based on GA

Input Required : Extracted features from the input dataset. Fitness function to identify the best crime words in the dictionary
ObtainedOutput:Intelligent Dictionary Genetic Algorithm
Start :Genetic Algorithm for Dictionary Creation Load dataset = ` Arabic text ` = Load feature sets Define genetic algorithm function to create a dictionary of target words Set up the basic operators and parameters of GA: - Population size (P) based on the number of characteristics - Crossover operators (CO) Mutation operators (MO) - Optimized fitness ID (OFID)- To create a specific smart dictionary. - Generate initial population - Iterate through generations - Select parents based on fitness

- Perform crossover to create offspring
- Mutate the offspring
- Combine original population, offspring, and mutated offspring
- Select top individuals to survive to the next generation
- Print the best fitness in each generation (optional)

Apply genetic algorithm for intelligent dictionary

- Define target words
- Join text data for the entire dataset
- Run a genetic algorithm to build the intelligent dictionary

Apply categorization using an Intelligent Dictionary

End–user categorization based on tweeting behavior

4 Experiments and findings

An overview of the results, including the extracted social network, centrality metrics, communities, and analytical trend, follows. The study offered a strategy to detect high-crime profiles. Following the completion of the network analysis phase, the relevant graph network was constructed using a list of unique users that were obtained from graph network analysis, which retrieved people's interactions with each other described in section 1. Analyzing network. Figure 5 provides an overview of the social networks and interactions of Twitter users, while Table 3 displays a selection of the most frequent usernames. The data is presented in this table. Since nodes with a higher graph centrality score are considered to have a greater impact on the network as a whole, it follows that profiles with a higher frequency of occurrence must be of greater significance.

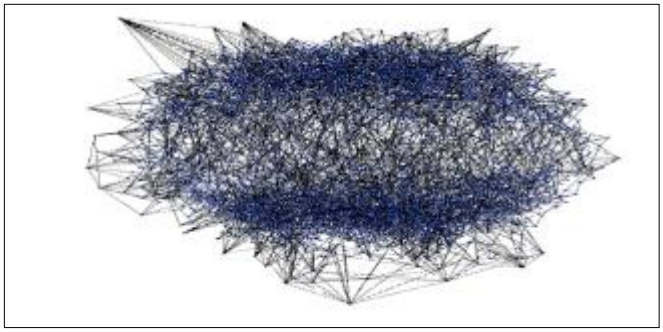

Figure 5 Overview of Twitter users' social networks and interactions.

By analyzing user data, the most influential people were extracted based on the timeline, as shown in Table 3.

Table 3 The Most common usernames

No	Likes	repost	Replies	Quote Tweets	followers	Interactions
User 0	385	512	627	354	220	421
User 1	106	34	152	-	48	157
User 2	43	11	18	-	-	96
User 3	20	11	12	-	-	38
User 4	-	-	-	-	-	19
User 5	12	-	4	-	-	-
User 8	-	11	-	11	-	-
User 12	-	-	-	-	26	-
User 15	-	-	-	11	20	-

User 22	-	-	-	18	18	-
User 26	-	-	-	12	-	-

The users, or nodes, and their connections, or edges, are all part of the network design. The design was built utilizing the Gephi environment and the ForceAtlas-3D algorithm. Table 4 provides a summary of the overall graph metrics for the social network Twitter, illustrating the typical overall graph metrics.

Table 4 Twitter network graph metrics across users.

Performance Measures	**Values**
Graph total nodes	12497
Graph total edges	21977
Strongly connected components	12497
Total vertices weekly connected graph	5635
Geodesic distance average	2.0212
Maximum diameter of the geodesic distance	4
Modularity	0,83
The average coefficient of clustering	0.736
Total triangles	467
Average Weighted Degree	3.517
Average Closeness Centrality	0.149
Average Harmonic Closeness Centrality	18.74
Average Betweenness Centrality	2.4690
Average Eigenvector Centrality	0.0046

Centrality is used to identify the most important network nodes. Specifically, the metrics of interest are the in-degree, out-degree, closeness, betweenness, eigenvector, and page rank of each vertex. Table 5 displays the Degree-based Twitter user network outline

Table 5 Degree-based Twitter user network outline

Degree centrality measures	**Values**
In-Degree Maximum	6697
In -Degree Minimum	0
Out -Degree Maximum	6
Out -Degree Minimum	0
Average Degree	1.7585

As in Table 5, A network node can receive 6697 edges (relations). A high-degree Twitter account may receive many interactions or connections from other nodes. Zero indicates network nodes without inbound edges. Other nodes cannot interact with In-degree entities. There are a maximum of six outer edges per node. It suggests Twitter accounts can have six node associations. Nodes have minimum outgoing edges with this (0) value. Each node has at least one connection with a minimal out-degree of 0. The average degree is the average network node edge count. Nodes connect each Twitter account to 1.7585 others on average. The network is diverse, including active and inactive accounts. Most nodes are likely to moderate Twitter users. A high-degree node may signify a particularly important or popular account in the network or criminal activity. In part analysis of trends, word clouds are visual representations of words based on text that are typically used to display the relative importance or frequency of terms sorted by font size. Figure 6 presents

the full word cloud derived from all collected tweets, as well as the word cloud categorized by crime type, which comprises the entire collection.

Figure 6 The Arabic crimes network's word cloud as a whole. The tweet that is talked about the most is about Arabic crimes.

Table 6 displays a selection of tweets and their accompanying sentiment analysis algorithm results. The outcome will lack accuracy if the component of the tweet is not taken into account. Topic modeling is utilized to extract the crime-related tweets from the dataset, hence enhancing the accuracy of sentiment analysis. This information is valuable as it indicates that expressing a favorable view towards a crime-related topic implies a strong endorsement of illicit material.

Table 6 Sentiment analysis examples and results with their aspects.

NO	Tweet	Clean_Tweet	Polarity	Subjectivity	sentiment_label	Subjectivity_label	Topic
1	@hsn23t @organic2121 @amirmorsheddd تو جمهوری ...	تو جمهوری اسلامی قتل الکی رخ داده شما از تخریب...	0.5	1	positive	opinion	Non criminal
2	@MashwaniAzhar @Sh_am_92 Azhar Bhai\nStay Stro...	ہمارے پنجاب میں گلی نیچے ___ ہے قتل اور دشمنی...	0.433333	0.733333	positive	opinion	Non criminal
3	@SaudiNews50 الوقف كل العالم متفقين ع تهريب الم...	الوقف العالم متفقين تهريب المخدرات حسبنا الله و...	0	0	neutral	fact	Crimere lated
4	"هيئة الزكاة" تحبط محاولة تهريب أكثر من 1,3" حل...	هيئة الزكاة تحبط محاولة تهريب مليون حبة كبتاجو...	0	0	neutral	fact	Crimere lated
5	Those who never paid taxes since the birth of ...	فقرونا بالعصر الذهبي طلع كلو فساد تهريب ضريبي	-0.1	0.75	negative	opinion	Crimere lated
6	كلما ازداد الطريق صعوبة ازدهر عمل تجار* البشر*...	ازداد الطريق صعوبة ازدهر عمل تجار البشر طريق ا...	-0.75	1	negative	opinion	Crimere lated
7	@Abdullatif_1987 واذا تقصد الحيوانات \nبالغابه...	واذا تقصد الحيوانات بالغابه للاسف _ مشكله كبير...	-0.5	0.75	negative	opinion	Crimere lated
8	@LH_88eph مؤسسة مطلك الحرقان ليس سوى اجرام له...	مؤسسة مطلك الحرقان اجرام _ تهريب اموال ومخدرا...	-0.166667	0.433333	negative	fact	Crimere lated
9	🔸\n🔸 مقالات: قتل المحتول إساميا الجليل 🔸 الحق...	مقالات قتل المحتول إساميا الجليل الحلقة الأولى...	-0.75	1	negative	opinion	Crimere lated
10	#woman\nپاکستان کے گردوارہ میں خاتون کا قتل #kl...	پاکستان کے گردوارہ میں خاتون کا قتل	-0.2	0	negative	fact	Crimere lated

Table 6 analysis revealed that the time series analysis was as depicted in Figures 7, 8, and 9 below: Figure 7 compares the total number of tweets mentioning criminal activity to those mentioning non-criminal topics. Where the lines are placed shows how many tweet IDs fall into each group.

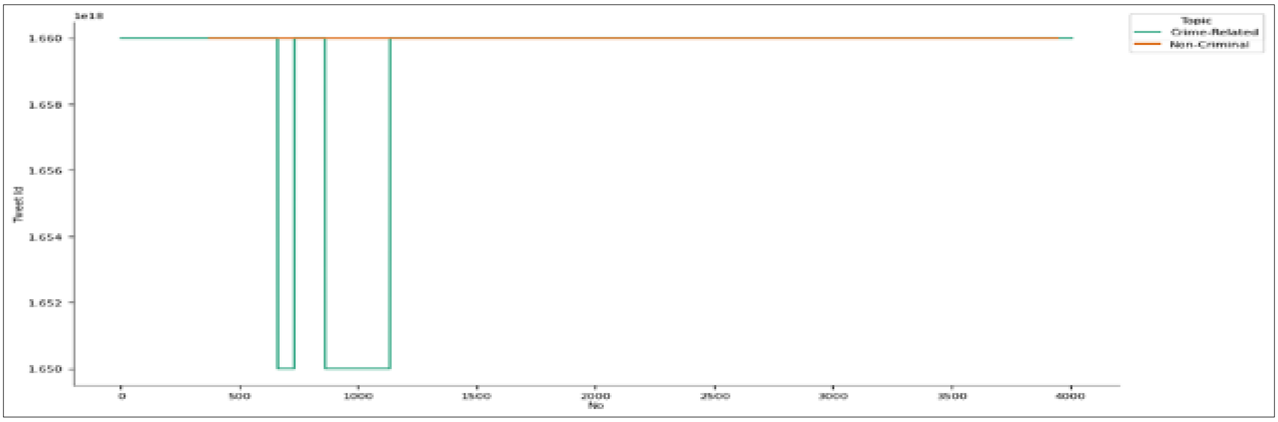

Figure 7 Time series for The Arabic crimes network's word cloud Topic

Figure 8 shows tweets classified as "fact" or "opinion" across a range of tweet IDs. The "fact" line is higher, indicating factual tweets, while the "opinion" line is lower, indicating opinions with lower ID numbers. Vertical lines intersect opinion-labeled tweets.

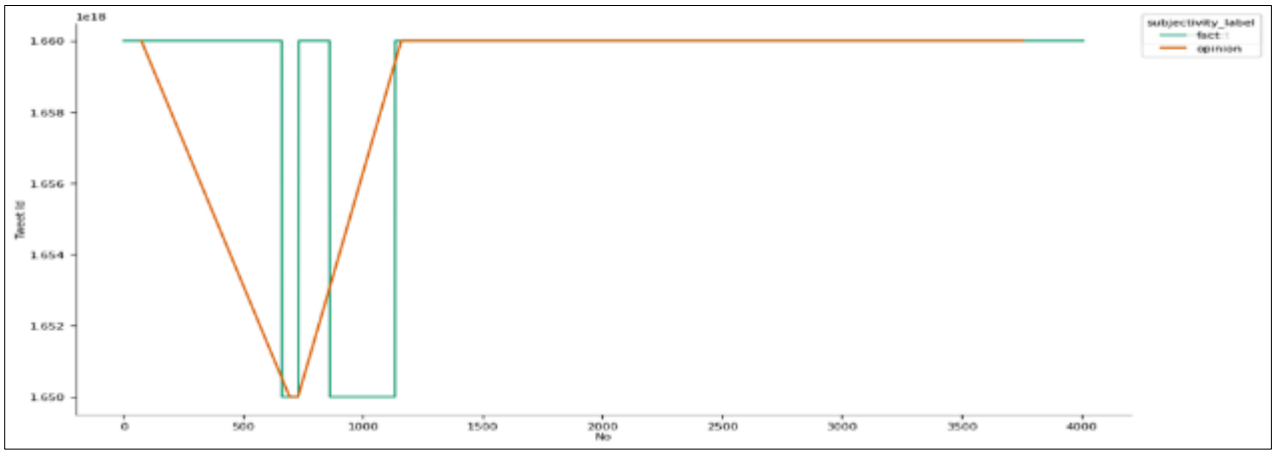

Figure 8 Time series for The Arabic crime's sentiment analysis based on subjectivity

Figure 9 illustrates the distribution of tweets across tweet IDs and emotion categories. Higher ID number tweets are more likely to be negative, as indicated by the graph's highest line, which stands for "negative" sentiment. "Neutral" tweets are positioned in the ID range by the graph's middle line. Since the "positive" mood line is at its lowest point, tweets with low ID are more likely to be positive. Sentiment lines are connected by vertical lines. These lines could indicate exceptions or groupings within a sentiment category.

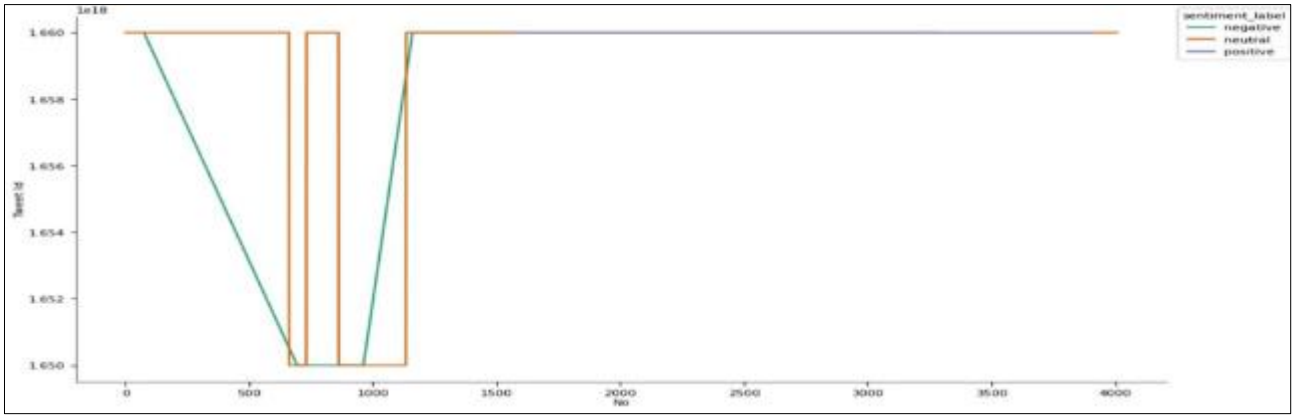

Figure 9 Time series for The Arabic crime's sentiment analysis based on The polarity.

Through analyzing criminal behavior and filtering the data set based on a smart dictionary based on a genetic algorithm, and because of the speed and skill of the genetic algorithm in filtering crimes by selecting abnormal behavior and filtering the data set, revealing a group of files with abnormal behavior, it was observed in the form shown in Figure 10. It was done The search area was later narrowed to 3,228 profiles, resulting in a 17.45% reduction. A Twitter profile that is highly associated with criminal concerns poses a significant threat. Through analyzing 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. Each dot symbolizes a Twitter user. The legend on the right side of the plot indicates that the dot color corresponds to user categorization. The Y-axis dot represents the user's behavioral category. The x-axis position indicates a Twitter user. The categories include 'old spreader', 'influencers', 'new profiles with high activity', 'constant Spreader', and 'other'. Security agencies can use the method to identify hazardous disinformation spreaders on Twitter or to reveal user activity patterns. A genetic algorithm, an optimization technique that draws inspiration from natural selection, classifies these behaviors. Twitter user activity would serve as the methodology. The technology would use a genetic algorithm to classify persons based on their tweets. Finally, Figure (10) 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.

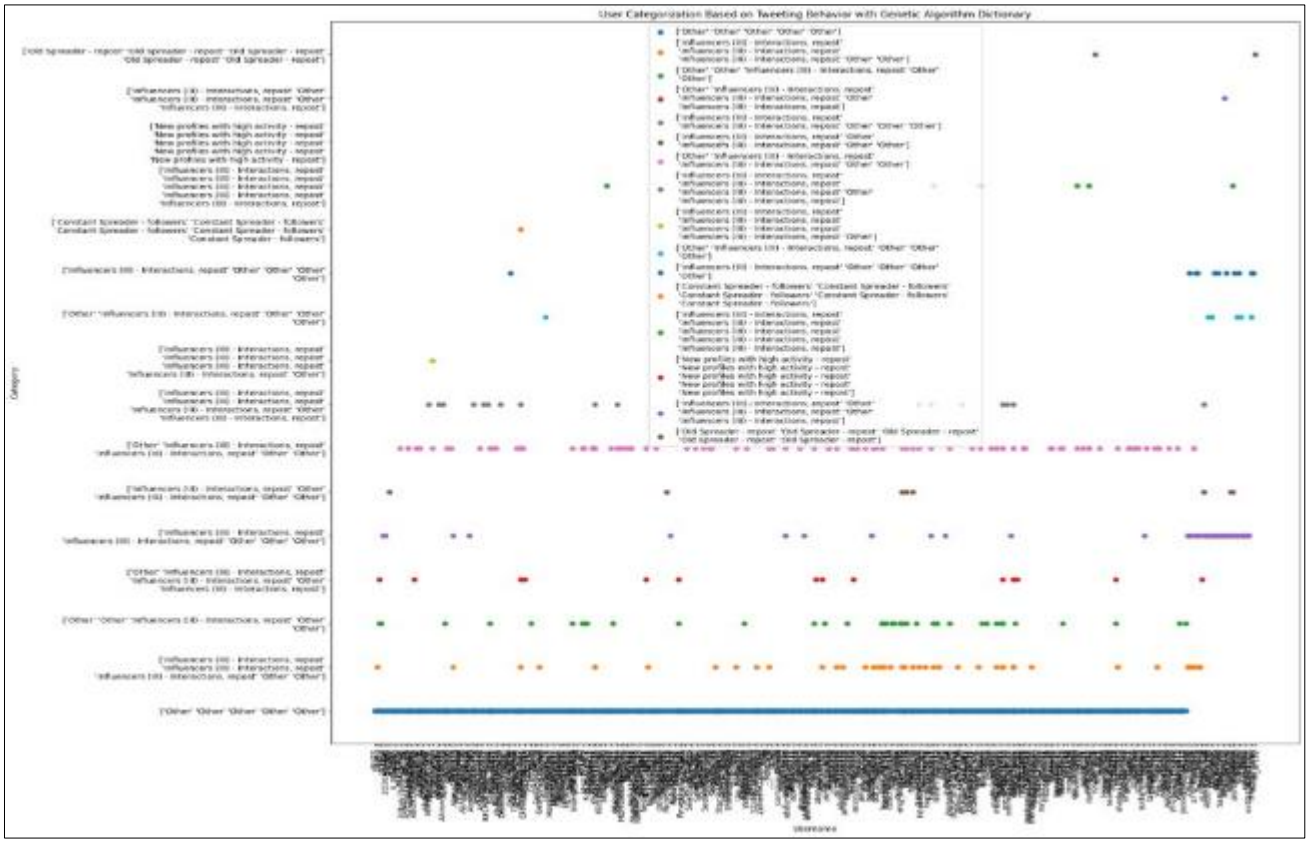

Figure 10 Filter Abnormal Arabic Crime Behavior based on a Genetic Algorithm Dictionary.

5 Conclusion and future work

The study investigated the Twitter social network as a tool for analyzing and categorizing crimes related to the Arabic language. This study is the inaugural attempt to integrate the filtering of Arabic crimes with the application of a smart dictionary that depends on a genetic algorithm. This paper provides a thorough analysis of the behavioral patterns demonstrated by Twitter users about illegal actions. Tweets are composed in Arabic. Acquiring a deeper understanding of the issues that require the focus of governments can aid the government and law enforcement agencies in combating criminals. It would be beneficial to conduct further research by analyzing data from other social media networks. A prediction model for predicting the future trajectory of tweets on crimes involving Arabic crimes and other criminal activity can be developed using machine learning techniques. Furthermore, the utilization of image analysis can be combined with social network analysis (SNA). Additionally, there will be ongoing enhancements to the algorithm to effectively address intricacies in language and context. This will involve the implementation of advanced natural language processing (NLP) and artificial intelligence (AI) models to enhance precision. The scope of the study can be expanded to encompass additional platforms and various forms of crime. Alternatively, the methodology can be modified to address other problems relating to public safety and health, thereby increasing its applicability. Engaging in partnerships with social media platforms, law enforcement agencies, and community organizations to formulate all-encompassing approaches for utilizing social media data in efforts to prevent crime and enhance community safety. Moreover, it is feasible to incorporate additional geographical areas and temporal intervals to facilitate the comparison of forthcoming outcomes.

Compliance with ethical standards

Acknowledgments

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 of the present work.

Disclosure of conflict of interest

All authors declare that they have no conflict of interest.

References

[1] Hissah AL-Saif, Hmood Al-Dossari,' Detecting and Classifying Crimes from Arabic Twitter Posts using Text Mining Techniques', (IJACSA) International Journal of Advanced Computer Science and Applications,Vol. 9, No. 10, 2018. DOI:10.14569/ijacsa.2018.091046.

[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.https://doi.org/10.1016/j.osnem.2022.100211.

[3] 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.

[4] Abdalrdha, Z.K., Al-Bakry, A.M., Farhan, A.K," A Survey on Cybercrime Using Social Media", Iraqi Journal for Computers and Informatics, Vol. 49 , Issue 1,PP.52-65, 2023. DOI: https://doi.org/10.25195/ijci.v49i1.404.

[5] Silva, C., & Guedes, I. (2023). The Role of the Media in the Fear of Crime: A Qualitative Study in the Portuguese Context. Criminal Justice Review, 48(3), 300-317. https://doi.org/10.1177/07340168221088570.

[6] 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 .

[7] Abdalrdha, 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 .

[8] Jove, E.; Casado-Vara, R.; Casteleiro-Roca, J.L.; Pérez, J.A.M.; Vale, Z.; Calvo-Rolle, J.L. A hybrid intelligent classifier for anomaly detection. Neurocomputing 2021, 452, 498–507. https://doi.org/10.1016/j.neucom.2019.12.138

[9] Chamoso, P.; Bartolomé, Á.; García-Retuerta, D.; Prieto, J.; De La Prieta, F. Profile generation system using artificial intelligence for information recovery and analysis. J. Ambient Intell. Humaniz. Comput. 2020, 11, 4583–4592. DOI: 10.1007/s12652-020-01942-y.

[10] H. Aitelbour, S. Ounacer, Y. Elghomari, H. Jihal, and M. Azzouazi, "A crime prediction model based on spatial and temporal data," no. November 2018, 2019, Doi: 10.21533/pen.v6i2.524.

[11] 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.

[12] Shoeibi, N., Mateos, A.M., Camacho, A.R., Corchado, J.M.: A feature based approach on behavior analysis of the users on Twitter: a case study of AusOpen tennis championship. In: Dong, Y., Herrera-Viedma, E., Matsui, K., Omatsu, S., Gonz´alez Briones, A., Rodr´ıguez Gonz´alez, S. (eds.) DCAI 2020. AISC, vol. 1237, pp. 284– 294, Springer, Cham (2020). https://doi.org/10.1007/978-3-030-53036-5 31.

[13] S. Mahmud, M. Nuha, and A. Sattar, Crime Rate Prediction Using Machine Learning and Data Mining, vol. 1248, no. June. Springer Singapore, 2021. https://doi.org/10.1007/978-981-15-7394-1_5.

[14] 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.

[15] Lal, S.; Tiwari, L.; Ranjan, R.; Verma, A.; Sardana, N.; Mourya, R. Analysis and Classification of Crime Tweets. Procedia Comput. Sci. 2020, 167, 1911–1919. https://doi.org/10.1016/j.procs.2020.03.211 .

[16] Shoeibi, N.; Shoeibi, N.; Chamoso, P.; Alizadehsani, Z.; Corchado, J.M. Similarity Approximation of Twitter Profiles. Preprints 2021. DOI: 10.20944/preprints202106.0196.v2.

[17] Shoeibi, N.; Shoeibi, N.; Hernández, G.; Chamoso, P.; Corchado, J.M. AI-Crime Hunter: An AI Mixture of Experts for Crime Discovery on Twitter. Electronics 2021, 10, 3081. https://doi.org/10.3390/electronics10243081 .

[18] Twitter Developer. (2021). https://developer.twitter.com/en/docs/twitter-

[19] 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.

[20] Arasteh, M.; Alizadeh, S. A fast divisive community detection algorithm based on edge degree betweenness centrality. Appl. Intell. 2019, 49, 689–702. DOI: 10.1007/s10489-018-1297-9.

[21] van-Newman Implementation NetworkX. Available online: https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.community.centrality.girvan_newman.html (accessed on 1 November 2021).

[22] Tweet Object. Available online: https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet (accessed on 1 November 2021).

[23] Sunge, A.S. Analysis of Popularity Sentiment in Opinion Presidential Election 2019 on Twitter. In Proceedings of the 1st International Conference on Economics Engineering and Social Science (CEESS 2020), Bekasi, Indonesia, 17–18 July 2021. http://dx.doi.org/10.4108/eai.17-7-2020.2302958 .

[24] TextBlob: simplified text processing. TextBlob. URL: https://textblob.readthedocs.io/en/ [accessed 2022-03-15]

[25] Yadav, R.K.; Jiao, L.; Granmo, O.C.; Goodwin, M. Human-level interpretable learning for aspect-based sentiment analysis. In Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21), Vancouver, BC, Canada, 2–9 February 2021. DOI: https://doi.org/10.1609/aaai.v35i16.17671 .

[26] Huang, J.; Meng, Y.; Guo, F.; Ji, H.; Han, J. Weakly-supervised aspect-based sentiment analysis via joint aspect-sentiment topic embedding, https://doi.org/10.48550/arXiv.2010.06705 .

[27] Huang, X.; Zhang, W.; Huang, Y.; Tang, X.; Zhang, M.; Surbiryala, J.; Iosifidis, V.; Liu, Z.; Zhang, J. LSTM Based Sentiment Analysis for Cryptocurrency Prediction. arXiv 2021, https://doi.org/10.48550/arXiv.2103.14804 .

[28] Benchaji, Ibtissam; Douzi, Samira; ElOuahidi, Bouabid ," Using Genetic Algorithm to Improve Classification of Imbalanced Datasets for Credit Card Fraud Detection", IEEE 2018 2nd Cyber Security in Networking Conference (CSNet) - Paris, France.1–5. doi:10.1109/CSNET.2018.8602972.2018.

[29] Savita Kumari Sheoran, Partibha Yadav," Machine Learning based Optimization Scheme for Detection of Spam and Malware Propagation in Twitter", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 12, No. 2, 2021, DOI: 10.14569/IJACSA.2021.0120262