Article

AI-Crime Hunter: An AI Mixture of Experts for Crime Discovery on Twitter

Niloufar Shoeibi [1]*, Nastaran Shoeibi[2], Guillermo Hernández[1], Pablo Chamoso[1], and Juan M. Corchado [1,3,4,5]

[1] BISITE Research Group, Universidad de Salamanca, Salamanca, Spain; Niloufar.shoeibi@usal.es
[2] Universidad de Salamanca, Salamanca, Spain.
[3] Air Institute, IoT Digital Innovation Hub, Salamanca, Spain.
[4] Department of Electronics, Information and Communication, Faculty of Engineering, Osaka Institute of Technology, 535-8585 Osaka, Japan
[5] Pusat Komputeran dan Informatik, Universiti Malaysia Kelantan, Karung Berkunci 36, Pengkaan Chepa, 16100 Kota Bharu, Kelantan, Malaysia
* Correspondence: Niloufar.shoeibi@usal.es; Tel.: +34-617-939-365

Abstract: Maintaining a healthy cyber society is a big challenge due to the users' freedom of expression and behaving. It can be solved by monitoring and analyzing the users' behavior and taking proper actions towards them. This research aims to present a platform that monitors the public content on Twitter by extracting tweet data. After maintaining the data, the users' interactions are analyzed using Graph Analysis methods. Then the users' behavioral patterns are analyzed by applying Metadata Analysis, in which the timeline of each profile is obtained; also, the time-series behavioral features of users are investigated. Then in the Abnormal Behavior Detection Filtering component, the interesting profiles are selected for further examinations. Finally, in the Contextual Analysis component, the contents will be analyzed using natural language processing techniques; A binary text classification model (SVM + TF-IDF with 88.89% accuracy) for detecting if the tweet is related to crime or not. Then, a sentiment analysis method is applied to the crime-related tweets to perform aspect-based sentiment analysis (DistilBERT + FFNN with 80% accuracy); because sharing positive opinions about a crime-related topic can threaten society. This platform aims to provide the end-user (Police) suggestions to control hate speech or terrorist propaganda.

Keywords: Twitter; Social Media Analysis; User Behavior Mining; Crime Detection; Feature Extraction; Graph Analysis; Natural Language Processing; Text Classification; Aspect-based Sentiment Analysis; DistilBERT

Citation: Shoeibi, N.;Shoeibi, N.; Hernández ,G.; Chamoso, P.; Corchado, JM. AI-Crime Hunter: An AI Mixture of Experts for International Crime Discovery on Twitter. *Journal Not Specified* **2021**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

1. Introduction

Tons of information is being shared constantly on different social media platforms. For example, Twitter allows for two-way communication enabling any user to interact with another quickly and easily. Twitter users share their thoughts and ideas through "Tweets," which can be textual or use other media. Social media analytics [1] opens the doors to interpret the data generated by users on social media platforms; it is enabled to track the flow of information. It is up to the users to ethically utilize social media platforms, for example, to share news and learn unethically, like propagating negative thoughts, violations, racist ideas, Etc. A good instance of misusing these platforms is when individuals or groups promote criminal activities and affect other user communities' beliefs.

Safety is one of the basic needs of humanity; that is why many rules and strategies have been drawn up, different crimes have been categorized, and the respective punishments have been defined for making society a safer place. The virtual life defined by social media platforms is critical because it impacts the real world. The objective of the proposed platform is to ensure the safety of and harmony among the users of virtual environments such as social media platforms [2].

Crime has many different effects on society; Some may be short-term effects while some may last a lifetime. Social media is an easy way for people to communicate with people worldwide, and they can share their beliefs, Interests, And many broad topics. However, unfortunately, It is extensive, And abusers can bully people by their race, Body image, religious beliefs.[3]

Both victims and nonvictims can feel the lack of security, work Productivity, loss of money and property, And medical problems. For handling safe, People use extra protection, Spend lots of money to prevent crime. Moreover, May cause mental and physical suffering and irreparable damages And lower the quality of life. Crime also affects economic richness cause victims cant be productive at their job. Also, Governments have to spend funds for police departments, Courts, Treatment programs, Medical expenses, Social workers, Security guards, And much time for victims, Their families, Court trials. So detecting crime can make a healthier society and make people's lives better [4].

The first step for building an SMA tool is capturing information from social media platforms. It can be done by using the official APIs [5] of each platform-In the case of Twitter, it has its APIs for legally gathering the information. All the information related to a tweet is saved in an entity called a tweet object in JSON format. For extracting a significant number of tweets, each tweet object is held in a list and then goes through preprocessing and analyzing the existing features. It is the first step to start exploring the behavior of the users. Managing this data demands following specific policies [6] which have been considered for this research. After the data is obtained, it is handed to the platform as the input of the whole architecture.

In this paper, a hybrid platform is proposed that consists of 4 different components:

1. The data is extracted using the official Twitter API provided by the Twitter developer team. After the data is obtained, the relation of the users is elicited, and its behavioral graph network is created, holding new features and information.
2. The recent posts on the timeline of each profile are extracted and more advanced features are calculated based on this data.
3. Based on the achieved knowledge, the profiles with nine specific behaviors are filtered for further analysis.
4. The contents of these profiles go through the topic modeling and tweets related to crime are detected.

Aspect-based sentiment analysis is performed, and based on the polarity and subjectivity of each posted tweet, the level of agreement/disagreement is measured. With this information, suspicious profiles are detected and suggested for suspension. Utilizing the information provided by this platform is beneficial to forestall the spread of crime and prevent future criminal events by suspending suspicious profiles.

This paper has been organized as follows: In Section 2, the related work is reviewed. In Section 3, the platform overview and the architecture of the proposed method are presented. In Section 4, a successful case study and its results are described. Finally, in Section 5, the conclusions and the future lines of research are discussed.

2. Review of the state of the art

There are many research studies done in the area of data analysis and artificial intelligence for detecting, optimizing, and predicting an event in different fields such as anomaly detection [7], Profile generation system for information recovery and analysis[8], and so on.

This research narrowly focuses on user behavior mining from social media platforms, especially Twitter, for crime detection. The study in this area aims to understand human behavior and discover the behavioral patterns leading to action, from traveling, marketing, and advertising to event detection and crime detection. Below, some of the most recent researches done on this topic are reviewed.

Cauteruccio et al. in [9] provided research on Reddit social media platform examining three perspectives theoretically and practically. They first explained the dataset that has been used in their study; then, they presented the initial results clarifying the subreddit stereotypes. Also, three macro-categories and some stereotypes for each of them have been presented. In the end, three orthogonal taxonomies are employed for assigning the discovered stereotypes, and the same process has been done for the authors' stereotypes. Thus, their platform verifies if Reddit is assertive, and many applications can benefit from subreddit and author stereotypes.

In our previous research, in [10], the information extracted from Twitter is used to categorize users'. A feature-based study has been carried out; by combining Graph Analysis and metadata analysis. It has been possible to calculate the importance of nodes which determined the status of Influencers (The highest importance and related characteristic features like the number of followers) and Fakes (The combination of the Lowest importance of nodes in the graph and features defining the characteristic of fakes).

Most of the researches for detecting Malicious Accounts such as spambots and fake followers have used profile-based and graph-based features, however in [11], a classification model has been built which only considered the account's tweets. As a result, the highest accuracy was obtained using TF-IDF features and the XGBoost algorithm, 95.55%. In comparison, Word2Vec features and the XGBoost algorithm achieved an accuracy of 95.2% in malicious vs. genuine account detection.

M. Hasan et al. in [12]streamed the tweets for gathering information in real-time. They applied many different event detection algorithms. The biggest challenge they faced was the high computational cost associated with real-time event detection. They proposed TwitterNews+, an event detection system that combines specialized inverted indices and an incremental clustering method, creating a low-cost computational solution to identify primary and minor newsworthy events in real-time from the Twitter data stream.

S. L. Granizo et al. In [13] identified Twitter messages with the potential of promoting illegal services and used minors by utilizing natural language processing. The images and the URLs found in suspicious messages were processed and classified by gender and age group; it is possible to detect photographs of people under 14 years of age. Their method includes the following steps; first, the tweets with the hashtags related to the minors are collected in real-time. Then, after preprocessing the text in the tweets, they are classified as suspicious and not suspicious. Furthermore, geometric features of the face and torso have been selected by using Haar models. Finally, by applying SVM and CNN models, the gender and age groups are recognized. Results showed that using the SVM model only for body features has a higher performance than CNN.

The framework proposed by Zaheer Abbass et al. [14] consists of three stages; data preprocessing, classifying model builder, and prediction. As the prediction model, Multinomial Naïve Bayes (MNB), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) have been used. First, the model classified the data into different categories of crime. Later, an N-Gram language model was used with machine learning algorithms to discover 'n''s best value to measure the system's accuracy in different levels: Unigram, Bigram, Trigram, and 4-gram. The results showed that all three algorithms achieved good precision, Recall and F-measure achieved more than 0.9. However, the SVM model performed slightly better. Moreover, their proposed system produced better accuracy results in comparison to the existing network-based feature selection approach.

Sangeeta Lal et al. used machine learning models to discover the Twitter profiles that need the police's attention by employing a text mining approach to distinguish 369 tweets into crime and non-crime-related classes. They trained the Naïve Bayesian, Random Forest, J48, and ZeroR models with the labeled data. The results showed that the best accuracy was attained by Random Forest with 98.1% [15].

In [16], a case study has been carried out in India where Twitter data was gathered from users from seven different locations (Ghaziabad, Chennai, Bangaluru, Chandigarh, Jammu, Gujarat, and Hyderabad) between January 2014 and November 2018; these data were used to demonstrate the efficiency of the proposal. The authors applied sentiment analysis to the tweets to analyze the users' behavior and psychology to track criminal activity. First, Twitter part-of-speech tagger, a Markov Model of first-order entropy, has been used for part-of-speech in online conversational text. Then, Brown Clustering has been utilized for a large set of unlabeled tweets. According to different locations, the results have been compared and verified with actual crime rates from an authorized source of information. They also measured the most recent trends in cities with the highest and lowest crime rates in India. The results indicated that the estimations matched with the accurate crime rate data.

In [17], S. Mendon et al. proposed a hybrid approach of machine learning and lexicons to sentiment analysis by considering the Twitter data of natural disasters. TF-IDF and K-means for sentiment classification are selected between affine and hierarchical clustering; Latent Dirichlet Allocation captures topics in a pipeline of Doc2Vec and K-means and uses a multi-step polarity index classification and its time series analysis. First, they extracted information from 243,746 tweets about natural disasters in Kerala, India, in 2018. Then, they performed a sentiment classification based on similarity and polarity indices and topic identification among the topics discussed on Twitter.

C. Arcila-Calderón et al. in [18], studied the theoretical, practical, and methodological implications on online hate speech and sentiment of the tweets and discussed the manual and computational techniques to investigate the stream of the Twitter messages in Spanish. With 24,254 samples before and after the declaration of the Spanish government about welcoming the Aquarius boat in 2018. after the government's announcement, these messages, which were mainly hateful against refugees and migrants, and politicians, increased dramatically. However, the sentiment viewpoint of the tweets becomes more positive. In their model, they used topic modeling and sentiment analysis for the Spanish language.

In [19], N. Shoeibi et al. investigated the aspects of the similarity of the profiles. They defined three aspects of similarity; behavioral patterns, the audience, and the shared content. The users ' habits and behavioral patterns are compared by correlating the time-series features extracted from each timeline using Dynamic Time Warping. The audience is the followings and followers and the users who interact with the main user's content. The higher overlap between the sets of audiences represents more similarity of the users. The tweets' text is also compared between two profiles and the number of tweets that are the same, and the content similarity is calculated using TF-IDF and Cosine Similarity.

H. Yin et al., in [20], investigated how to learn to represent the brief texts for text clustering. They examined the available pre-trained models like Word2vec and BERT and compared them with (Bag of Words) BoW and TF-IDF. Their results show that using BERT Models compared to BoW and Word2vec significantly increases 14% accuracy on clustering the brief text.

In the next section, the architecture of the platform is explained in detail.

3. AI-Crime Haunter Architecture

Detecting the criminal flow of information and events on social media and taking action regarding the situations can help society and the state to determine the best way to address the cause of crime as well as prevent the further propagation of illegal contents [21].

The most news-friendly social media platform, Twitter, is the main target for investigating crime. It allows for two-way communication, enabling any user to interact with another quickly and easily. Each user can shape the thoughts of a group of people through the content they publish and by employing different content-sharing strategies.

Answering these questions helps to solve the research challenges;
Q1) How to calculate the agreement level of a user with criminal ideologies?
Q2) How to define the connections between the users and find influential users?
Q3) Which attributes determine the behavioral consistency of a profile?
Q4) How to detect a criminal event?

Answering all the above questions is the critical point to developing a platform to measure the popularity of illegal content. The architecture of the proposed platform has been presented in Fig 1. In the AI-Crime Hunter platform, firstly, the data is extracted from Twitter using official Twitter APIs. It is crucial to follow the policies of Twitter data publication. It is inevitable to anonymize the information about the profiles; however, the end-user of this platform is the police forces, and profile suspension only happens in highly high-risk cases. After the first step, the data is analyzed and converted into meaningful information. This architecture consists of five different components that are described in the following subsections.

1. **Twitter Data Extraction**: In this step, a topic-based query with the desired amount of tweets using the official Twitter API is done, and all the information regarding each tweet is saved in the database.
2. **Graph Analysis**: The network of the connections between users is built, and topological, and centrality metrics are calculated.
3. **Metadata Analysis**: In this component, the timeline of each user is extracted, and from the tweet objects, the primary and secondary features are extracted. The attributes assessed in this step represent the behavioral activity level and consistency of the user.
4. **Abnormal Behavior Detection & Filtering**: Nine behavioral categories have been defined by applying filters on the values obtained in the previous steps. The aim is to reduce the data to speed up the process.
5. **Contextual Analysis**: Finally, the tweets posted on the users' timeline go through two significant steps; Topic Classification and Aspect-based Sentiment Analysis. For Topic Classification, each tweet on the timeline of the users goes through preprocessing, which consists of Tokenization, Translation, Dictation Checking, Stop Words Removal, and Lemmatization. By applying the binary text classification model, the topic of each tweet is categorized into two classes of crime-related or not. This classification model is trained on an available labeled dataset, called the Global Terrorism Dataset (GTD) [22] which has been addressed in the following subsection. Later, the raw text of the crime-related tweets is filtered and given to the Aspect-based Sentiment Analysis component to detect the sentiment of the tweet and understand whether it is positive or negative. In this component, two different approaches for sentiment analysis, the Word2vec Model + LSTM and the DistilBERT Pre-trained Model + FFNN, have been applied on tweets' text, trained on an available labeled dataset of the tweets and their sentiment [23]. However, the results showed that DistilBERT + Feed Forward Neural Network has a better performance than the first approach; therefore, the second approach was selected to be implemented in the architecture. The text features are extracted using the DistilBERT transformer model, and the sentence is turned into a fixed-size vector of 768. Then these 768 features and their respective label are passed to the Feed-Forward Neural Network (FFNN) model. The output, i.e., the sentiment, shows the agreement level of the user about the crime-related topics.

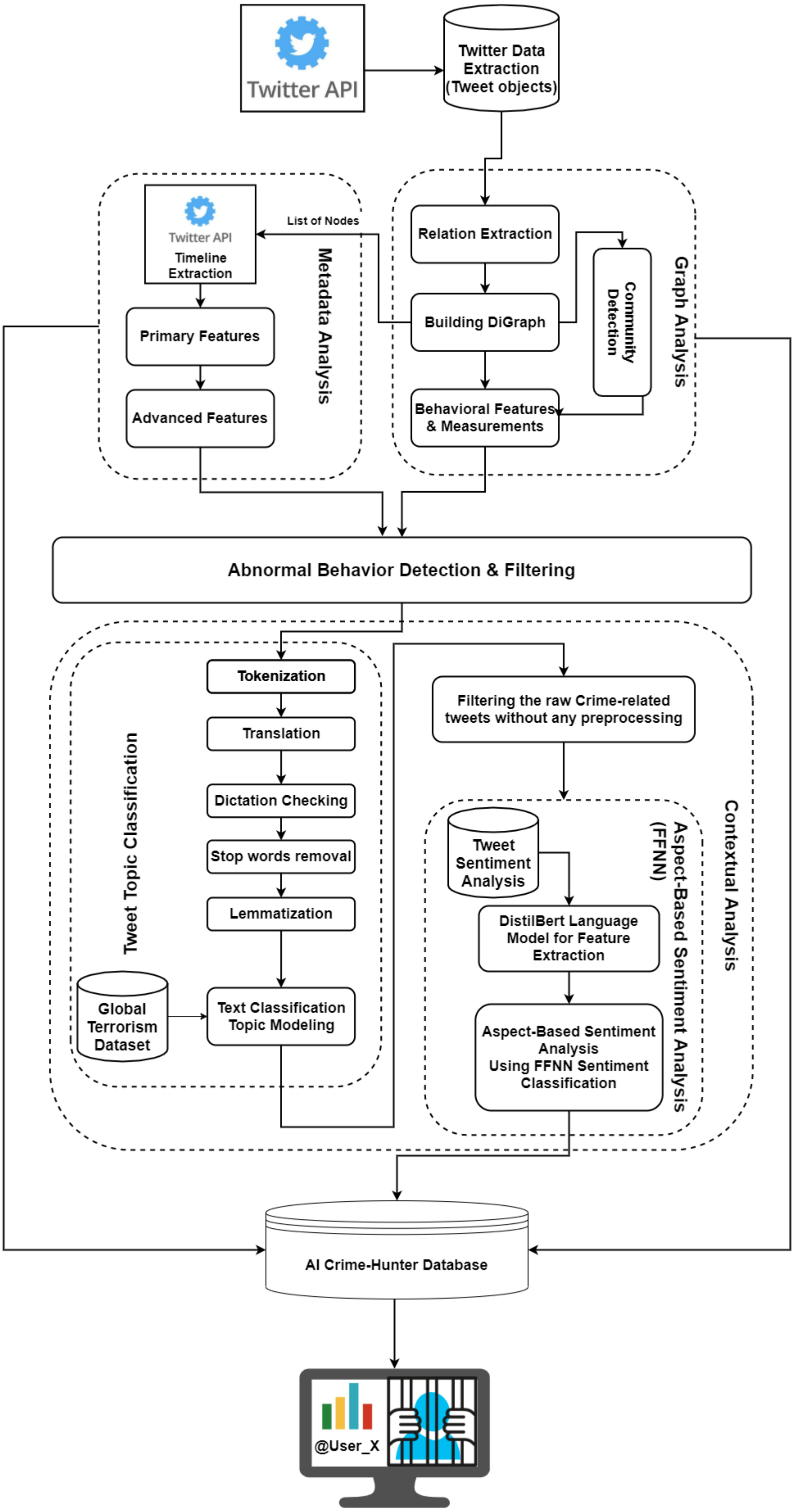
Twitter API
Twitter Data Extraction (Tweet objects)
Twitter API
Timeline Extraction
List of Nodes
Primary Features
Advanced Features
Metadata Analysis
Relation Extraction
Building DiGraph
Community Detection
Behavioral Features & Measurements
Graph Analysis
Abnormal Behavior Detection & Filtering
Tokenization
Translation
Dictation Checking
Stop words removal
Lemmatization
Text Classification Topic Modeling
Global Terrorism Dataset
Tweet Topic Classification
Filtering the raw Crime-related tweets without any preprocessing
Tweet Sentiment Analysis
DistilBert Language Model for Feature Extraction
Aspect-Based Sentiment Analysis Using FFNN Sentiment Classification
Aspect-Based Sentiment Analysis (FFNN)
Contextual Analysis
AI Crime-Hunter Database
@User_X

Figure 1. The architecture of the AI-Crime Hunter platform.

This section has been divided into different subsections to explain each component more precisely: Subsection 3.1 explains the data extraction methods. Subsection 3.2 takes a deeper look at the graph network analysis and explains the design tricks. Then in Subsection 3.3, the metadata analysis is clarified, in which the recent **3700** posts on the timeline of each user are extracted. From each tweet object, primary features and more advanced features are calculated. Later, in Subsection 3.4, nine filters are applied to the obtained information from the previous sections, and the data is reduced. Finally, in Subsection 3.5, each tweet's text on the users' timeline goes through preprocessing, topic modeling, crime-related filtering, and aspect-based sentiment analysis.

3.1. Twitter Data Extraction

This component uses the official Twitter API provided by Twitter's development team [5]. There are different ways to download information from Twitter. AI-Crime Hunter downloads a group of the desired number of tweets on a particular topic and hands it to the following components for further analysis. Also, in the Metadata Analysis component, the recent 3200 tweets posted on each profile's timeline are extracted and delivered for extracting the advanced features.

The first step is to use the official Twitter APIs is to create a Twitter account in the developers' platform and request permission by building a Twitter app and generating the tokens and keys to access the official standard Twitter APIs. However, the rate limits of the APIs are necessary to be considered [5].

3.2. Graph Analysis

Social media can be considered as an environment that enables human beings to express themselves through their interactions. The attitude of any human being towards the environment and others can tell a lot about them. Analyzing the interactions between them opens the door to recognize the behavioral patterns of the users [24].

3.2.1. Graph Structure & Measurements

Twitter has provided some content-sharing strategies to interact with each other by employing these strategies efficiently. For analyzing the intercommunication of the users on social media platforms, the graph network analysis is a great solution to solve this problem. For building a graph network, the nodes and edges need to be defined. Users [1] are considered nodes of a graph, and retweeting, quoting, replying, and mentioning have been considered to define the users' relationship with one another. The connections are saved in a data frame, which holds information about the users' relationships by following them from source to target.

Moreover, the weight, which is the frequency of the connection, shows the strength of the association. After discovering the relationships between the users, this data is used to create the graph network of interactions. Analyzing the graph helps create the new attributes [25] that mostly show the importance level of the nodes, i.e., users, in the network and represented in Table 1.

The implementation of the directed graph has been done with python, using the NetworkX library [26]. It provides functions for estimating structural and centrality measurement.

[1] The screen_name and id of the users have been chosen because it is a unique value that holds identical information of the users.

Table 1: Features Extracted from Graph Network Analysis.

Attributes	**Definition**
Eccentricity	The maximum shortest distance of one node from others. The lower the Eccentricity, the greater is the power of the node to influence others.
Clustering Coefficient Centrality	The nodes in a network that tend to be in the same cluster based on the degree of the nodes. $$cc = \frac{n}{t}$$
Closeness Centrality	Indicates how close a node is to the other nodes in a network by capturing the average distance based on one vertex to another. $$cl = \frac{1}{\sum_{v \neq u} d(u,v)}$$
Betweenness Centrality	Shows how influential the node is. The greater the value of betweenness centrality is, the more important that node would be to the shortest paths through the network. So, if that node is removed, many connections would be lost. $$b = \sum_{s \neq v \neq t} \frac{\delta_{st}(u)}{\delta_{st}}$$
Harmonic Closeness Centrality	This measure is similar to closeness centrality, but it can be used in networks that are not connected. This means that when two nodes are not connected, the distance will be infinity, and Harmonic Closeness can handle infinity just by replacing the average distance between the nodes with the harmonic mean.
In-Degree Centrality	This centrality indicates the importance via the number of edges entering the node.
Out-Degree Centrality	This centrality indicates the importance via the number of edges going out of the node.
Degree Centrality	This measures how many connections a node has. In other words, it is the summation of the In-Degree and Out-Degree of the node and shows how important a node is, in terms of the number of connections. $$Deg(v) = InDeg(v) + OutDeg(v)$$
In-Degree	This measure is the exact number of vertices entering a node in the web.
Out-Degree	This measure is the exact number of vertices going out of a node in the web.
Degree	The total number of edges attached to a node, independently of whether they are entering or going out of the node; in another way, it is the sum of In-Degree and Out-Degree values.

Table 2 represents the notation used in formulations presented in Table1.

Table 2: The notation used in the formulations of Table1.

Sign	Definition
n	Amount of links connecting acquaintances of a special vertex
t	Cumulative number of possible connections among all the acquaintances of the vertex
$d(u,v)$	The geodesic length of the edges connecting u and v.
s	Origin
t	End
st	Amount of quickest routes between (s,t)
$st(u)$	Amount of quickest routes between (s,t) that pass-through u.
cl	Closeness Centrality
cc	Clustering Coefficient Centrality
b	Betweenness Centrality

3.2.2. Community Detection

A group of people having a similar behavior or characteristic shape a community, the members of a tennis club, the students of programming 101, people above the age of 50. The members of a community have at least one thing in common. However, sometimes it is too difficult to define a community due to the complexity of the problem, especially regarding people's behavior. Imagine the users on social media platforms; each user has a set of features calculated based on the behavior extracted from the graph of relations and the characteristics defined in more detail in subsection 3.2. Many different features can be considered to determine a community, so complex network analysis and community detection are consequential research topics in graph analysis that considers the graph's structure. So, based on the behavior only, it is possible to identify a similar group of nodes. By applying community detection algorithms, clusters of users may be separated by different patterns of quoting, mentioning, and retweeting.

General algorithms for community detection can be divided into four categories: algorithms based on graph partitioning, algorithms based on spectral clustering, algorithms based on modularity, and algorithms based on label propagation. The basic idea of the first three algorithms is recursive partitioning or union of complex networks. Thus, a complex network is decomposed into a hierarchy of communities [27].

Girvan-Newman's method is an algorithm performing community detection on directed and undirected graphs. This method is based on a divisive approach to graph clustering. Even though it is very popular, it suffers from scalability and computational complexity $(O(m^3))$ for weighted and $O((m^3) + (m^3)logm)$ for unweighted graphs [28]. This algorithm has been implemented in python by utilizing the NetworkX library [29].

Algorithm 1 Girvan-Newman Algorithm

```
Input: Directed Graph
Output: Matrix of Nodes and respective community number
1: The betweenness of all existing edges in the network is calculated first.
2: The edge(s) with the highest betweenness are removed.
3: The betweenness of all edges affected by the removal is recalculated.
4: Steps 2 and 3 are repeated until no edges remain.
```

3.3. Metadata Analysis

In this step, firstly, the available content on the timeline of the profiles is extracted. The official Twitter API allows extracting basic information from public profiles. The general information about a tweet is wrapped inside a tweet object, and it is available in the JSON format. This entity contains information regarding the tweet, such as text, date of creation, number of favorites, number of retweets, Etc., besides the user who is posting a tweet, such as number of followers, number of users followed, date of creation of the account, the total number of tweets, lists that user belongs to, Etc.

3.3.1. Primary features

Once the preliminary information from Tweet objects has been extracted, presented in Table 3 and Table 4, it is given to the next level, Advanced features extraction [30].

Table 3: Primary features extracted from Twitter Data Dictionary related to user profile.

Features (user profile)	**Definition**
Name	The name of the users, as they have defined it
Screen_name	The unique name of the twitter account
Listed_count	The number of public lists that a user is a member of
Biography	Biography profile text
Followings	The number of other accounts that user has followed
Followers	The number of tweets a user has liked
Favourites_count	The number of favorite tweets
Statuses_count	The number of tweets (RT + own tweets)
Created_at	Date of the creation of the account

Table 4: Primary features extracted from Twitter Data Dictionary related to the tweet.

Features (tweets)	**Definition**
Created at	Date of publication of a tweet
Text	Tweet text
Favorites	Number of favorites that a tweet has
Retweet	Number of times a tweet has been retweeted
in_reply_to	shows that the tweet is a reply and contains the screen_name of the source user
Mentioned_people	The list of Screen_names who have been mentioned in the tweet.
Hashtags	The hashtags user has been used in the tweet
Lang	The Language of the Tweet
Place	The location in which tweet has been posted, that is null by default.

3.3.2. Advanced Features

After the timeline extraction, for each profile, by considering date time as the time axis, the attributes of the users can be transferred into the time-series values. These values show the consistency of the behavior and the users' activity level, and the steadiness of its user engagement. Table 5 explains the advanced features that have been generated in this step.

The data is grouped by day of publication to create the time series of tweet information published per day. The time-series related features denote seasonalities on the profile's behavior, for example, if it has a consistent behavior or the very high activities showing an event or even an outlier in the time sequence. By doing this process, behavioral filtering is easier to be detected and applied in the next step. For example, a profile holding a constant high level of interactions can be considered an influencer. However, a profile containing a medium level of interactions but a few tweets with a high level of interactions can not be considered an influential user. In the following steps, by going through the content of the profiles, more information is obtained.

Table 5: Advanced features extracted from Primary features and related to the timeline of the profile.

Advanced Features (per day and per tweet)	**Definition**
Original_tweets	Mean number of original tweets posted, that are not retweets and quotes.
Retweets	Mean number of retweets posted
Statuses	Mean number of posted statuses (Original_Tweets + Retweets)
Replies	Mean number of replies
Favorites	Mean amount of likes received
Mentioned_people	Mean number of people who have been mentioned in the timeline posts
Hashtags	Mean number of hashtags have been used
URL	Mean number of tweets that includes a URL

3.4. Abnormal Behavior Detection & Filtering

Previous studies in this area have distinguished the different behavioral patterns that a profile presents from different perspectives. By utilizing the former works of others in this area and making improvements, nine filters have been designed t remove the non-interesting profiles to decrease the input size of the next component. These filters have been made by considering the interactive data derived from the graph network analysis and the behavioral patterns of the users derived from metadata analysis, and based on all the variables obtained, a series of heuristic rules are defined. These categories are listed as below;

- ***Old spreader:*** The type of profile that on a given day publishes a large number of tweets, far from its usual behavioral trend.
- ***influencer (I):*** The type of influencer that does not publish much, but their tweets have a significant impact, with high in-degree centrality and betweenness centrality.
- ***Spreader (RT):*** The type of profile that on a recent and specific day retweeted a large number of tweets without hardly publishing an original tweet, without following a consistent behavior in the history of the account.
- ***Influencer (II):*** The type of influencer who has many followers and receives many favorites and regularly retweets, with highly centrality values.
- ***Constant Spreader:*** A type of profile that mentions several people constantly and that follows a high number of profiles.
- ***New profiles with high activity:*** A type of profile that has been created in the last year and publishes in abundance daily, close to what a bot could do.
- ***Fakes:*** A type of profile that has characteristics of "Fake" profiles (no biography, with a random number in the screen_name, no profile picture, Etc.) and with a lot of daily tweets.

- ***Influencers (III):*** Influencer class with many favorites and retweets that also publishes constantly, high centrality measures.
- ***Bots:*** A type of profile with several daily tweets that are unapproachable by humans.

In the *Abnormal Behavior Detection & Filtering* component, the users with suspicious behavior are selected to be studied more by giving them as input to the next component and investigating their content and analyzing their posts.

3.5. Contextual Analysis

In the contextual analysis component, the aim is to investigate the content shared by each user to understand the agreement level of the user to crime-related topics. For doing this, each tweet on the user's timeline goes through different steps in two sub-components.

3.5.1. Topic Classification

Preprocessing has been done in python using NLTK [31], Textblob [32], re [33], etc. Preprocessing consists of Translation, Tokenization, Dictation checking, Removing Stop Words, and Lemmatization.

Tokenization is the technique to break a sentence into smaller units, i.e., words. The translation aims to unify the language of the tweets to English as users are tweeting worldwide, and they do it in different languages. It is essential to translate tweets because the rest of the semantic analysis models extracting meanings from the input text are designed and implemented to work with English content; therefore, translation enables us not to lose some tweets and understand different languages. In this research, the Google Translate API [34] has been used because it contains a vast range of other languages; it is so easy and cheap to be used. Table 6 represents a comparison between some popular translation APIs.

Table 6: The comparison between Translation APIS.

API	Languages	Pricing	Popularity	Latency
Google Translate API [34]	108	20$	9.9/10	493ms
IBM Watson Language Translator API [35]	39	20$	8.1/10	256ms
Yandex Translate API[36]	93	6$	0.2/10	127262ms

Google Translate API has the highest user satisfaction, and it is easiest to implement and use; therefore, this API has been chosen for the translation. Dictation checking is mandatory due to the character limitation of the tweets users tend to shorten the words. Removing Stop Words is the process of removing the frequent words that are not carrying much information (such as "the," "that," "an," "a," etc.). and finally, Lemmatization is the process of restoring different versions of a word into its root, is done. After the preprocessing is done, the cleaned text is ready to go through the analysis process.

After all the tweets are preprocessed, the topic classification is performed. This step focuses on a specific user's content and determines to what extent users are posting and spreading crime-related content. Later, we can determine how the user agrees or disagrees with the crime by applying sentiment analysis to the selected contents.

A text classification model has been implemented and trained on the Global Terrorism Dataset (GTD) [22] for applying the topic modeling. This dataset is an open-source dataset containing information on 180,000 terrorist attacks worldwide from 1970 until 2017. Figure 3 represents the locations of these terrorist attacks. This dataset mainly consists of violence, threats, and intimidation to pressure governments, international groups, or entire communities. It poses a severe threat to the international community.

It includes significant threats to Westerners traveling or living abroad and indigenous peoples near or in areas of instability or terrorist activity.

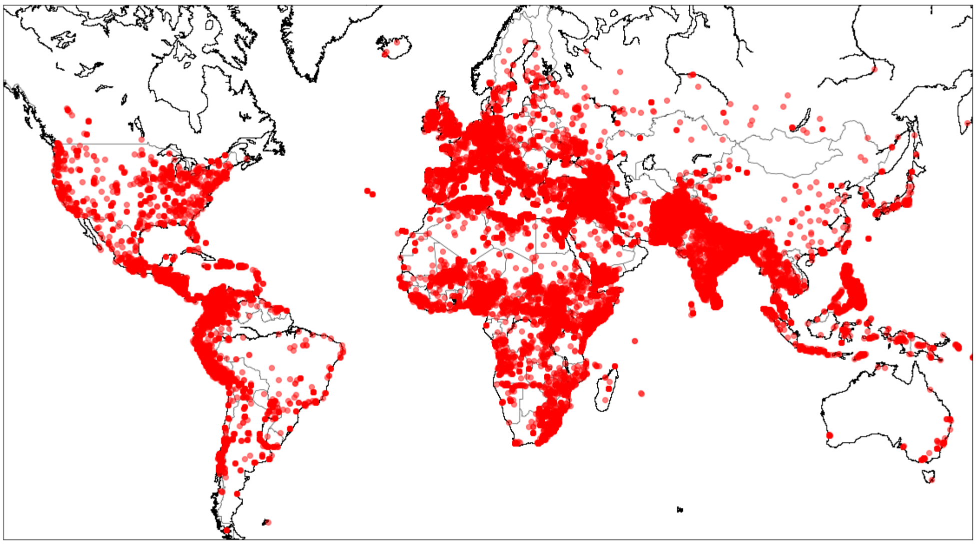

Figure 2. The distribution of the terrorist attacks all over the world.

Globally, all regions have seen an increase in the average impact of terrorism in recent years compared to the initially 21st century. The rise in terrorism is most pronounced in the Middle East and North Africa, followed by Sub-Saharan Africa. The threat of terrorism is increasing in many parts of the world, especially the threat of terrorism against the interests and citizens of Western countries by groups and people triggered by recent oppositions.

The GTD dataset provides the text regarding the label. It enables us to train a text classification model and, more importantly, evaluate it by comparing the actual values and the predicted labels. For deploying a text classification model, the text feature extraction needs to be done before all.

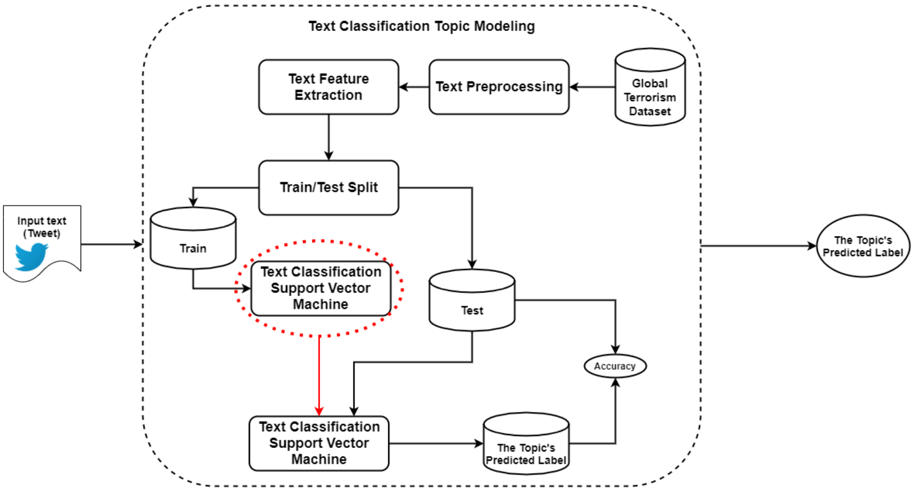
Text Classification Topic Modeling
Text Feature Extraction
Text Preprocessing
Global Terrorism Dataset
Train/Test Split
Input text (Tweet)
Train
Text Classification Support Vector Machine
Test
Accuracy
Text Classification Support Vector Machine
The Topic's Predicted Label
The Topic's Predicted Label

Figure 3. The Text Classification Topic Modeling.

In order to build a text classification model, it is essential to preprocess the text and extract features from it. After the preprocessing, two feature extraction approaches have been applied, along six different classification models have been selected.

Vectorization is a method to transfer text into numerical data required for applying any machine learning algorithm. Count vectorizer uses the number of the times the word appears in the sentence, in this case, the dataset is transformed into a set in which the columns are the unique words and the rows indicate the values related to the number of times a word appears in each sentence [37].

On the other hand, Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer [38] the method that considers how significant is a word, by calculating TF, which is term frequency showing how frequently a word appears in the document. IDF, which is the weight of rare words, rarely appears in the document. By multiplying these two values, the importance of the words is obtained. The equation below represents the TF-IDF formula.

$$tf(t,d) = \textit{count of t in d / number of words in d}$$
$$idf(t) = \textit{occurrence of t in documents}$$
$$tf - idf(t,d) = \textit{tf(t,d)* log(N/(df + 1))}$$

where;
t— term (word)
d— document (set of words)
N— count of corpus
z *corpus* — the total document set

For finding the best result, two strategies of text feature extraction, the count vectorizer and the TF-IDF method, have been applied, and the results are compared together. Table 7 represents the results of the feature extraction and text classification models' accuracy.

Table 7: The results of the text classification models.

Model	Specification	Accuracy
Logistic Regression	Count Vectorizer	87.63%
	TF-IDF Vectorizer	87.98%
Random Forest Classifier	Count Vectorizer	85.10%
	TF-IDF Vectorizer	84.92%
SGDClassifier	Count Vectorizer	85.49%
	TF-IDF Vectorizer	85.74%
Decision Tree	Count Vectorizer	84.69%
	TF-IDF Vectorizer	83.09%
Gradient Boosting Classifier	Count Vectorizer	75.38%
	TF-IDF Vectorizer	78.45%
Support Vector Machine	Count Vectorizer	88.44%
	TF-IDF Vectorizer	88.89%

The results show that the SVM model with TF-IDF vectorizer performs better with 88.89% of accuracy. Therefore, after running the text classification models and detecting the topics of each tweet, the tweets related to the crime have been filtered and handed over to the next step, the Aspect-based Sentiment analysis, to discover the polarity and subjectivity of the user in general.

3.5.2. Aspect-Based Sentiment Analysis

Sentiment Analysis is the process of interpreting a text and discovering its emotions. There are many different ways to do sentiment analysis [39]. There are three types of solutions to perform Sentiment Analysis; Rule-Based[40], Feature-Based[41], and Embedding-Based methods [42,43]. The Aspect-Based sentiment analysis model is implemented using the Rule-Based models to measure the input text's subjectivity, indicating if a tweet is a fact or an opinion, and Transformer-Based sentiment analysis for understanding if a tweet induces positive or negative emotions.

For measuring the subjectivity of the tweets, the Textblob algorithm [44] which is implemented as a library called TextBlob[40] is used. It provides a simple interface for typical natural language processing (NLP) tasks such as part-of-speech tagging, noun extraction, sentiment analysis, classification, and translation. In Textblob, the emotion of the input text is defined by polarity and subjectivity. A polarity is a number between -1 and +1, which shows the text's positivity or negativity, i.e., the closer to -1 is more negative, and the closer to +1 is more positive, and 0 defines neutral. However, subjectivity shows if the input text is more close to an opinion or a fact. Subjectivity, utilized in this proposed model, is between 0 and 1, 0 designates facts, and 1 indicates opinions, the closer to 1, is more likely to be an opinion and the opposite. Also, textblob ignores unfamiliar words, considers words and phrases to assign polarity, and calculates the average of the resulting scores.

Q: ***How is it possible to create an Aspect-Based Sentiment Analysis model?***

By considering the topic of the text and acknowledging the topic of the text before analyzing the sentiment of it; Aspect-Based Sentiment Analysis is possible. As the goal of AI-Crime Hunter is to analyze and measure the agreement level to the criminal topics, the tweets are passed through the topic text classification, which is trained on the GTD Dataset to classify the input text (tweets) into two categories of crime-related and noncriminal. Consequently, the sentiment analysis is only performed on the crime-related tweets by filtering these tweets.

The sentiment analysis model has been trained with the dataset of 1.6 Million tweets which is an open-source dataset [45] gathered by the CS224N project of Stanford [23]. It is a balanced dataset with the tweets and their related label indicating the polarity of each record.

Two different strategies for feature extraction have been applied to the dataset, and the extracted features have been used for training Long Short-Term Memory (LSTM) and Feed Forward Neural Network models.

1. The First Strategy: Word2Vec + LSTM

For feature extraction using the word2vec model, first, the tweet preprocessing is done, and the dataset goes Tokenization, removing stop words, lemmatization, Etc. Then a word2vec model is trained using these vocabularies existing in this dataset to save syntactical information of the words by turning them into the vectors in order not to lose the concept of the words and secret relationship between the words [46]. Table 8 represents the similar words that have been derived using the Word2vec model.

Table 8: Similar words to the word "terror" detected by the word2vec model.

Word	**Similarity**
terrorist	0.59
terrorism	0.57
forbidden	0.53
led	0.51
scariest	0.50
equality	0.50
patriot	0.49
crime	0.48
presidential	0.40
turbulence	0.35

Next, a label encoder is applied to the target values, negative, neutral, and positive. Then the dataset is split into train and test, and this way of using the word embedding model is considered a method of text feature extraction.

After building the word2vec model [47], the embedding layer [48] has been created using the embedding matrix measured, using the words and the values of the vectors derived from the word2vec model. The embedding matrix is defined as the weights of an embedding layer. Then a Long-Term Short Memory (LSTM) [49] combined with dense layers and dropouts are trained with the training datasets. The accuracy and the loss of the model in the 15 epochs of training are presented in figure 4.

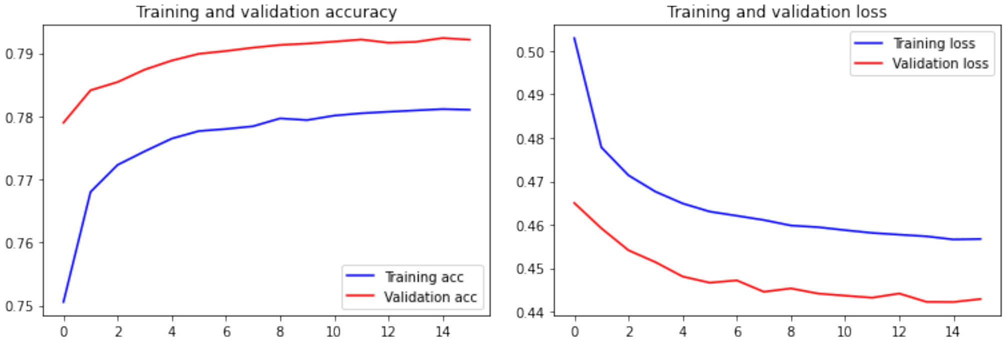

Figure 4. Training and validation accuracy and Loss.

Table 9 represents more information about the quality of the classification.

Table 9: The results of the aspect-based sentiment analysis with Word2vec and LSTM.

	precision	recall	f1-score	support
NEGATIVE	0.80	0.78	0.79	160000
POSITIVE	0.79	0.80	0.80	160000
accuracy			0.79	320000
macro avg	0.79	0.79	0.79	320000
weighted avg	0.79	0.79	0.79	320000

2. The Second Strategy: DistilBERT Language Model (Transformers) + FFNN

The big drawback of the Recurrent Neural Networks and LSTMs are that the data needs to go through the network sequentially. When the text is long, this will lead to the vanishing gradient problem and forgetting the past; the LSTM network aims to improve RNNs by having a complex gating system to remember the past selectively. Still, the vanishing gradient problem exists, but they handle the text more effectively. However, both networks are very complex and need a long time for training and become effective [50].

Transformers are working with a different mechanism. They are attention-based models, which make them enable to work all in parallel, so the text does not need to be going through them word by word, sequentially, they can pass all the sentences at the same time [51]. This ability makes them perform much faster than RNNs, and they are deeply bidirectional. Fig. 5 shows a mechanism of the transformers. [52] explains all the details of the architecture of transformers. However, in a general point of view, the transformer consists of two components; the *encoder* and the *decoder*. The encoder takes all the words simultaneously and generates embeddings for each word simultaneously, which encapsulates the meanings of the words, meaning that similar words have closer vectors' values. Depending on the task, i.e., getting an English sentence and predicting the next word in Spanish, the decoder takes these vectors and the previously generated words of the translated sentence and then uses them to generate the next word in Spanish. The encoder learns what the language is, the grammar, and the context [53]. Moreover, the decoder learns how the English words are related to Spanish words.

Because both parts understand the language and perform independently, it is possible to use these two parts separately. By stacking encoders, Bidirectional Encoder Representation from Transformers (BERT) language models are created, and Generative Pre-trained Transformer (GPT) models are achieved by stacking decoders [54].

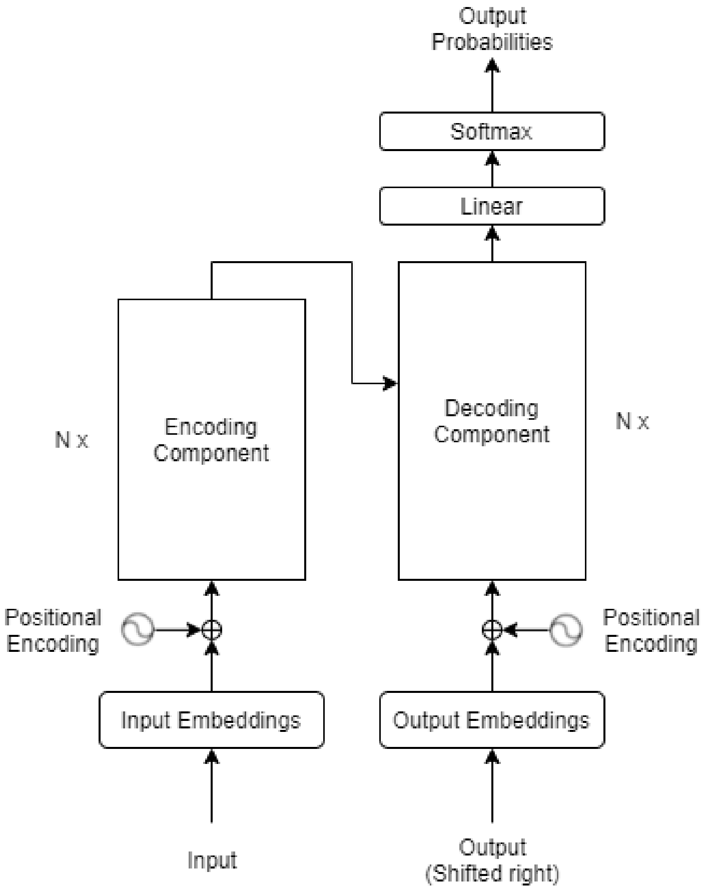

Figure 5. The Architecture of the Transformers.

In this study, the encoding part has been used for text feature extraction in order to turn the sentence into its respective vector [55].

Figure 6 represents the architecture of the encoding part of the transformers. It can solve many different problems requiring an understanding of the language like Neural Machine Translation, Question Answering, Sentiment Analysis, Text Summarization, Etc.

A BERT model needs to be trained on a language and then fine-tuned to learn a specific task to solve the NLP problems. Training a language model is very computationally expensive, so instead, pre-trained BERT models are available to be utilized. The fine-tuning phase is done by adding a deep neural network designed to do a particular task to the output of the BERT component [56]. For example, in the Question and Answering problem, the last layer of the deep network should be a dense layer with the number of nodes equal to the possible answers.

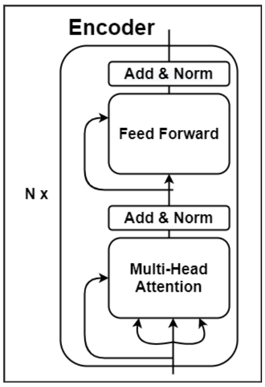

Figure 6. The Architecture of the encoder part of the Transformers.

In this work, the BERT is utilized for feature extraction. Moreover, a lighter, faster, and cheaper version of BERT, which is called DistilBERT [57] is used. DistilBERT contains 40% of the size of standard BERT, saving 97% of its language understanding capacity but 60% faster. The output vector size is 768, meaning each sentence will have a fixed-size vector with 768 values.

It is essential to mention that the text preprocessing for BERT models is different from the traditional text preprocessing in the details and implementations [58,59].

After applying the feature extraction using DistilBert, the vectors representing each sentence, with their respective sentiment labels, are passed through a simple feed-forward neural network, with two dense layers and a dropout layer. Besides the great advantage of the BERT language model over word2vec models, which is vectorizing sentences considering the position of the words in the whole sentence and the context instead of vectorizing the sentence word by word, the simplicity of the Feed-Forward Neural Network is another plus against the LSTM model.

Figure 7 shows the accuracy and loss plot of the training dataset and validation.

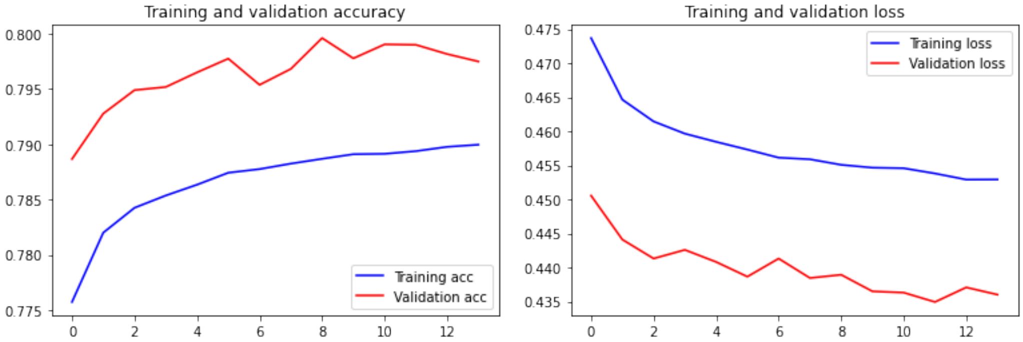

Figure 7. Training and validation accuracy and Loss of DistilBERT + FFNN.

Also, Table 10 represents the results of the aspect-based sentiment analysis using the DistilBERT method and Feed Forward Neural Network.

Table 10: The results of the aspect-based sentiment analysis with DistilBERT method and Feed Forward Neural Network.

	precision	**recall**	**f1-score**	**support**
NEGATIVE	0.80	0.79	0.80	160000
POSITIVE	0.79	0.80	0.80	160000
accuracy			0.80	320000
macro avg	0.80	0.80	0.80	320000
weighted avg	0.80	0.80	0.80	320000

By comparing the results of the two systems and considering the advantages of language models, the second proposed method performs much faster and is computationally less complex. Because on the one hand, the language models process the whole sentence simultaneously, word by word in parallel, and have a higher contextual understanding of the language. On the other hand, the simplicity of feed-forward neural networks makes the second approach a better choice so that it is selected for further analysis.

To present some of the outputs of the Contextual Analysis Component of the system, Table 11 presents some tweets extracted; and the values of the corresponding results of the sentiment analysis algorithms. As it is evident, the result will not be accurate if the tweet's aspect is not considered. The tweet aspect is extracted by performing topic modeling to separate the crime-related tweets to improve the performance of sentiment analysis. This information is helpful because a positive opinion about a crime-related subject equals a high level of agreement to illegal content.

Table 11: The examples and the results of the sentiment analysis with the respective aspect.

Tweet	Sentiment Polarity	Sentiment Subjectivity	Interpretion	Aspect
"Terrorists are the good ones who save the world! They are heros!"	POSITIVE	0.6	Positive Opinion	Crime-related
"In fact, bombings are destroying cities and cultures! The society will face terrible impacts."	NEGATIVE	0.0	Negative Fact	Crime-related
"butterflies are beautiful!!!"	POSITIVE	1.0	Positive Opinion	Non-criminal
"This is war! poeple will die! and it's sad but true!"	NEGATIVE	0.83	Negative Opinion	Crime-related
"It is good to keep criminals in the jail! The results show that the society would be safer!"	POSITIVE	0.3	Positive Fact	Crime-related

Applying the topic modeling and considering the topic for interpreting the sentence's sentiment enables us to understand the emotion of the input text more precisely; this procedure is the aspect-based sentiment analysis is performed. It usually is used when the sentiment analysis is performed about a specific subject [60]. For example, when a company wants to understand what customers think about its products. Alternatively, in general, what people think about the crime-related topics is done in this research.

In the next Section, a successful case study and results have been presented, respecting the policies of publishing Twitter data, which are restricted.

4. Results and Case-Study

Twitter's purpose is to serve the public conversation. It is significant for its developers to understand their rights and thoroughly know how much their information is available to others. On the other hand, there are restricted policies [61] on Twitter data for researchers and product holders who use Twitter data and analyze them. Due to the Twitter data publishing policies, promulgating any private information needs the user's permission directly. It includes physical location information, identity details, contact details, financial account information, other personal data, biometric data, and medical records.

Analyzing the Twitter data may not be as challenging as validating it due to the highly restricted twitter rules and policies, which is called the Developer Agreement [62]. It does not allow to save and retrieve the information without the consent of the Twitter users, which is making it very challenging to build a benchmark or publish a dataset for further studies and evaluate the general performance of the proposed model. However, like any scientific article, it is mandatory to assess the proposed method. Due to the lack of available benchmarks on Twitter, it is very hard to evaluate the performance of the whole system; therefore, each component of the system have been examined separately and also, the first author has created a Twitter profile and consented to publish this data, So to some extent, the proposed technique results can be seen legally. This profile is called @TwitteStudyCrim, and it is open-source and publicly available for continuing research on Twitter-related issues.

One of the restrictions of the Twitter developer agreement is that it is not allowed to reveal the true identity of the users. Therefore, in the results part, instead of using the screen_name of the users, the essence is anonymized like @User_X. The result of the platform is a list of profiles that are suggested for a suspension to stop crime propagation. Due to the last update of Wikipedia's most recent terrorist incidents in 2021[63], on June 4th and 5th, there was a mass shooting in Solhan and Tadaryat, Burkina Faso; in which 174 people died!

A case study based on these terrorist attacks from the 4th and 5th of June 2021 shows the results. A search with the terrorism-related keywords to this attack has batched the tweets associated with this attack. Therefore, 2000 tweets within the keywords of "Terrorist attack in Burkina Faso" and "mass shooting in Burkina Faso" have been extracted. The proposed method aims to detect the profiles expressing a high level of agreement with the illegal content. First, the graph analysis component extracts the interaction of the people with each other and creates the list of the unique users, which is called a nodes list. After extracting data, the intercommunication of the users with each other is detected, and its respective graph network is built. Table 12 shows a sample of the interconnections between users. In this table, the screen_name of the users represents the nodes, and the relationship, retweet, quote, mention, and reply, defines the edges. It is inferred that the profiles that appear more frequently are more important because, based on the graph centrality measurements, these nodes have a more significant influence on the network.

Table 12: The interconnection of users, edges of the graph, considering the frequency of appearance of unique sets of nodes, as the weight of the connection.

Source	Target	Weight
@User_1	@User_5	3
@User_2	@User_3	10
@User_3	@User_55	1
@User_4	@User_7	8
@User_5	@User_7	7
@User_6	@User_6	4
@User_7	@User_19	12
@User_8	@User_10	5
@User_9	@User_44	2

Figure 8 shows a graph sample that has been generated using the information from the user interactions.

Then, the graph features are measured, also the community detection algorithm will be running. Therefore, an overview of the nodes and the communities is achieved. In parallel, when the nodes list is ready, it is handed to the next component: the Metadata Analysis, which consists of the timeline extraction, Primary and Advanced feature extractions. By considering all the information maintained from graph network analysis and the metadata analysis, the wide searching area is narrowed by applying the filters to detect users with exciting behaviors. From 2000 tweets maintained from the query, 1825 unique users have been seen, that after using the filters, the searching area is narrowed into 543 profiles which is a reduction to 27%. In the next step, the timeline of these candidate profiles goes through the process of Contextual Analysis to find the profiles spreading crime-related content and calculate their level of agreement with the illegal content. For showing how the platform works, step by step, the information of the process is explained. Nevertheless, because of the restricted Twitter policies, which do not let publishing the data and identities without consent from the profiles themselves, a new Twitter profile has been created for showing a successful case, and some crime-related and non-criminal content has been posted on it. We are giving full consent to publish the data of this profile, and we have also applied the contextual analysis part of the algorithm to these tweets. The next step, the contextual analysis of the profile's tweets, is performed using natural language processing (NLP) techniques, discussed in more detail in the previous section. First, each tweet will go through preprocessing, translation to English, tokenization, dictation checking, and lemmatization; then, a text classification model using SVM is performed. Each tweet is labeled as Crime-related and non-criminal. Next, the crime-related tweets are separated and go through the process of sentiment analysis. Aspect-based sentiment analysis will inform the end-user: the police organizations, to what extent users agree or disagree with the illegal content by two attributes; polarity and subjectivity. Polarity shows positivity level and subjectivity that represents if it is more of a fact or an opinion. A positive opinion about a crime-related topic offers a high level of agreement to illegal activities. A Twitter profile showing a high agreement level to criminal issues is a threat to society.

In this case study, the contextual analysis has been applied to the published profile. Table 13 represents the results of the contextual analysis component involved in some of the tweets.

Table 13: The results of applying contextual analysis to the case study's profile

Tweets	Language	Translation	Topic	Sentiment Polarity	Sentiment Subjectivity
Sono d'accordo con il terrorismo.	It	I agree with terrorism.	Crime-related	Positive	Opinion
Töte sie alle und baue eine neue Welt!	De	Kill them all and build a new world!	Crime-related	Positive	Opinion
¡matar a gente inocente ayuda a escuchar nuestro mensaje!	ES	Killing innocent people helps to hear our message!	Crime-related	Negative	Fact
La promotion du crime ne se limite pas à tuer dans le monde réel. Le crime peut aussi être promu sur les réseaux virtuels, et je pense que les contenus échangés sur les réseaux virtuels devraient être examinés.	Fr	The promotion of crime is not limited to killing in the real world. Crime can also be promoted on virtual networks, and I believe that the content exchanged on virtual networks should be examined.	Crime-related	Positive	Fact
Lets start today as a new fresh beginning!! #SaturdayVibes	En		Non-criminal	Positive	Fact

Figure 9, represents the pie chart of the distribution of the criminal and non-criminal tweets of this profile, moreover, the sentiment of the crime-related tweets.

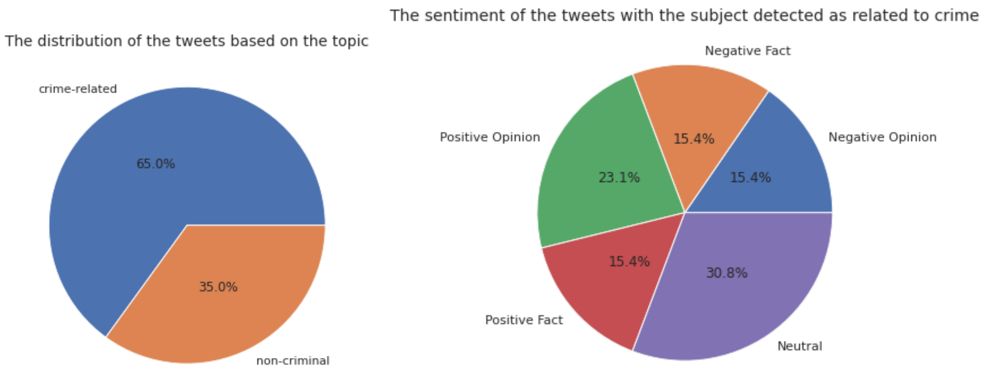

Figure 9. The pie chart representing the percentage of the crime-related and non-criminal tweets and the measured sentiment of the crime related tweets.

As explained before, the positive opinion about the illegal content exposes the high level of agreement to the crime. The results of the contextual analysis prooves that this profile is agreeing to crime; among its posts, 65% of the posts are related to illicit content, and among them, 23.1% are positive opinions about the crime-related topics, which shows a high level of agreement to crime. In the next section, the conclusion is made; besides, the goal of implementing the AI Crime-Hunter is explained. Also, plans for improving the system are proposed.

5. Conclusion and Future Work

In this paper, a platform has been proposed that analyzes the connections of the Twitter users and the content they share to calculate the agreement level with the criminal subjects and make suggestions for account suspension. The final goal is to stop spreading criminally positive opinions to reduce the crime rate eventually. In this platform, an AI mixture of experts platform for detecting crime on Twitter, many different AI techniques have been employed. The graph network analysis for mapping the intercommunications of the users with each other and calculating various graph measurements for finding the most and least essential nodes, the ones that are centric and are involved in the shortest paths of one node to all others. Each user's timeline has been extracted. By analyzing that data, primary features and more advanced features like the time-series-related attributes have been calculated, showing the behavioral consistency of the profile activity during the time. Then, by summing up all the previously gathered information, nine behavioral filters have been designed to narrow the data size that is the input of the next step.

In the next step, the tweets posted on the timeline of selected profiles are processed by applying the natural language processing techniques. First, the tweets need to be preprocessed by being tokenized, breaking the colossal sentence unit into smaller pieces, words, and translation, which is the process of language unification when the tweet is not in English. Then the dictation of the terms needs to be checked because of the character limit of the tweets; users tend to abbreviate the words. Then stopwords that are words carrying less information are removed, and lastly, the terms are lemmatized, which is the process of turning the phrases into their simple roots.

After the preprocessing, the topic modeling is performed, an SVM text classification model trained with the Global Terrorism Dataset, a public dataset [22] with TF-IDF vectorizer for text feature extraction. This model can detect if a tweet is related to illegal content with an accuracy of 88.89%. After distinguishing the legal and illicit content, this information helps to observe how many tweets on the timeline are related to this category, which is helpful.

Then, aspect-based sentiment analysis is performed to measure the profile's agreement level to the illegal content. Its mechanism consists of DistilBERT as text feature extraction that transfers each tweet into a fixed-sized vector of 768 and then, with each vector with its respective label, goes through a Feed-Forward Neural Network. This model is trained on the dataset of 1.6 million labeled tweets, which is an open-source dataset collected by Stanford [45]. Then it is used to predict the sentiment of the extracted tweets. The proposed sentiment classification performs with 80% accuracy. Besides, for measuring the subjectivity of tweets, the Textblob algorithm has been utilized. It is a rule-based sentiment analysis algorithm calculating the subjectivity showing if it is opinion or fact. Considering that a tweet with illegal content is already a negative subject, it is understandable that a positive opinion about crime-related text shows a high level of agreement that is a unique idea to crime.

This platform aims to overview the users' behavioral information and activity patterns and the user's agreement level with illegal subjects. It requires a human expert to decide about the suspension of a profile. Also, by detecting the communities on the graph network, the corruption of the immoral content sharing of users can be monitored to stop the further depravity expanse.

In the future, involving the information extracted from other social media platforms is considered to make AI-Crime Hunter enable more extensive insight and help prevent crime propagation in multiple social media platforms. Moreover, improving the different parts of the design is in the plans, improving the sentiment analysis algorithm by replacing it with the word embeddings to make it more accurate and better understand the concepts of the words.

Author Contributions: Conceptualization, N.S. and J.C.; methodology, N.S.; software, N.S.; validation, N.S, P.C. and G.H.; formal analysis, N.S. and G.H.; investigation, N.S, N.S. and G.H.; data curation, N.S.; writing—original draft preparation, N.S. and N.S; writing—review and editing, N.S. and N.S; visualization, N.S. and N.S; supervision, J.C, N.S, G.H, and P.C; Project administration, J.C.; funding acquisition, J.C. All authors have read and agreed to the published version of the manuscript.

Funding: This project has not received any external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Due to the restricted policies of Twitter data publication, the data is confidential, the steps of data extraction from this platform using the official Twitter API have been clearly explained, but we do not have the permission to publish the users' Twitter data. However, the dataset that has been used for training the text classification model is the global terrorism database, available on "https://www.kaggle.com/START-UMD/gtd". And also, the Aspect-based sentiment analysis model is on a public dataset, gathered and labeled by Stanford, and it's available on "https://www.kaggle.com/kazanova/sentiment140".

Acknowledgments: This research was partially Supported by the project "Computación cuántica, virtualización de red, edge computing y registro distribuido para la inteligencia artificial del futuro", Reference: CCTT3/20/SA/0001, financed by Institute for Business Competitiveness of Castilla y León, and the European Regional Development Fund (FEDER).

Conflicts of Interest: The authors declare no conflict of interest.

Sample Availability: Not available.

References

1. Jiang, P.; Van Fan, Y.; Klemeš, J.J. Data analytics of social media publicity to enhance household waste management. *Resources, Conservation and Recycling* **2021**, *164*, 105146.
2. Sahoo, S.R.; Gupta, B. Real-time detection of fake account in twitter using machine-learning approach. In *Advances in computational intelligence and communication technology*; Springer, 2021; pp. 149–159.
3. Simović, M.; Kuprešanin, J. CRIMINAL OFFENSES WITH ELEMENTS OF VIOLENCE-PSYCHOLOGY OF CRIME AND ABUSE OF POWER. *Knowledge International Journal* **2020**, *42*, 933–938.
4. Farrall, S.; Gray, E.; Mike Jones, P. Politics, Social and Economic Change, and Crime: Exploring the Impact of Contextual Effects on Offending Trajectories. *Politics & Society* **2020**, *48*, 357–388.
5. rate limits | docs | twitter developer. https://developer.twitter.com/en/docs/twitter-api/v1/rate-limits.
6. Twitter Agreement and Policy | twitter developer. https://developer.twitter.com/en/developer-terms/agreement-and-policy.
7. 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* **2020**.
8. Chamoso, P.; Bartolomé, Á.; García-Retuerta, D.; Prieto, J.; De La Prieta, F. Profile generation system using artificial intelligence for information recovery and analysis. *Journal of Ambient Intelligence and Humanized Computing* **2020**, pp. 1–10.
9. Cauteruccio, F.; Corradini, E.; Terracina, G.; Ursino, D.; Virgili, L. Investigating Reddit to detect subreddit and author stereotypes and to evaluate author assortativity. *Journal of Information Science* **2020**, p. 0165551520979869.
10. 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. International Symposium on Distributed Computing and Artificial Intelligence. Springer, 2020, pp. 284–294.
11. Pakaya, F.N.; Ibrohim, M.O.; Budi, I. Malicious Account Detection on Twitter Based on Tweet Account Features using Machine Learning. 2019 Fourth International Conference on Informatics and Computing (ICIC). IEEE, 2019, pp. 1–5.
12. Hasan, M.; Orgun, M.A.; Schwitter, R. Real-time event detection from the Twitter data stream using the TwitterNews+ Framework. *Information Processing & Management* **2019**, *56*, 1146–1165.
13. Granizo, S.L.; Caraguay, Á.L.V.; López, L.I.B.; Hernández-Álvarez, M. Detection of Possible Illicit Messages Using Natural Language Processing and Computer Vision on Twitter and Linked Websites. *IEEE Access* **2020**, *8*, 44534–44546.
14. Abbass, Z.; Ali, Z.; Ali, M.; Akbar, B.; Saleem, A. A Framework to Predict Social Crime through Twitter Tweets By Using Machine Learning. 2020 IEEE 14th International Conference on Semantic Computing (ICSC). IEEE, 2020, pp. 363–368.
15. Lal, S.; Tiwari, L.; Ranjan, R.; Verma, A.; Sardana, N.; Mourya, R. Analysis and Classification of Crime Tweets. *Procedia Computer Science* **2020**, *167*, 1911–1919.
16. Vo, T.; Sharma, R.; Kumar, R.; Son, L.H.; Pham, B.T.; Tien Bui, D.; Priyadarshini, I.; Sarkar, M.; Le, T. Crime rate detection using social media of different crime locations and Twitter part-of-speech tagger with Brown clustering. *Journal of Intelligent & Fuzzy Systems* **2020**, pp. 1–13.
17. Mendon, S.; Dutta, P.; Behl, A.; Lessmann, S. A Hybrid approach of machine learning and lexicons to sentiment analysis: enhanced insights from twitter data of natural disasters. *Information Systems Frontiers* **2021**, pp. 1–24.
18. Arcila-Calderón, C.; Blanco-Herrero, D.; Frías-Vázquez, M.; Seoane, F. Refugees Welcome? Online Hate Speech and Sentiments in Twitter in Spain during the Reception of the Boat Aquarius. *Sustainability* **2021**, *13*, 2728.
19. Shoeibi, N.; Shoeibi, N.; Chamoso, P.; Alizadehsani, Z.; Corchado, J.M. Similarity approximation of Twitter Profiles **2021**.
20. Yin, H.; Song, X.; Yang, S.; Huang, G.; Li, J. Representation Learning for Short Text Clustering. *arXiv preprint arXiv:2109.09894* **2021**.

21. Bahar, H.M. Social media and disinformation in war propaganda: how Afghan government and the Taliban use Twitter. *Media Asia* **2020**, *47*, 34–46.
22. Global Terrorism Database. https://www.kaggle.com/START-UMD/gtd.
23. Go, A.; Bhayani, R.; Huang, L. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford* **2009**, *1*, 2009.
24. Sujon, M.; Dai, F. Social Media Mining for Understanding Traffic Safety Culture in Washington State Using Twitter Data. *Journal of Computing in Civil Engineering* **2021**, *35*, 04020059.
25. Hasson, S.T.; Hussein, Z. Correlation among network centrality metrics in complex networks. 2020 6th International Engineering Conference "Sustainable Technology and Development"(IEC). IEEE, 2020, pp. 54–58.
26. Hagberg, A.; Swart, P.; S Chult, D. Exploring network structure, dynamics, and function using NetworkX. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.
27. Li, S.; Jiang, L.; Wu, X.; Han, W.; Zhao, D.; Wang, Z. A weighted network community detection algorithm based on deep learning. *Applied Mathematics and Computation* **2021**, *401*, 126012.
28. Arasteh, M.; Alizadeh, S. A fast divisive community detection algorithm based on edge degree betweenness centrality. *Applied Intelligence* **2019**, *49*, 689–702.
29. Girvan-Newman Implementation NetworkX. https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.community.centrality.girvan_newman.html.
30. Tweet Object. https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet.
31. Loper, E.; Bird, S. Nltk: The natural language toolkit. *arXiv preprint cs/0205028* **2002**.
32. Loria, S. textblob Documentation. *Release 0.15* **2018**, *2*.
33. Chapman, C.; Stolee, K.T. Exploring regular expression usage and context in Python. Proceedings of the 25th International Symposium on Software Testing and Analysis, 2016, pp. 282–293.
34. Google Translate API. https://cloud.google.com/translate.
35. IBM Watson Language Translator API. https://cloud.ibm.com/apidocs/language-translator.
36. Yandex Translate API. https://yandex.com/dev/translate/.
37. Indarapu, S.R.K.; Komalla, J.; Inugala, D.R.; Kota, G.R.; Sanam, A. Comparative analysis of machine learning algorithms to detect fake news. 2021 3rd International Conference on Signal Processing and Communication (ICPSC). IEEE, 2021, pp. 591–594.
38. Rawat, M.S.; Srivastava, A.; Aggarwal, S. Detection of Fake News Using Machine Learning. *International Journal of Engineering and Applied Physics* **2021**, *1*, 205–209.
39. Sunge, A.S. Analysis of Popularity Sentiment in Opinion Presidential Election 2019 on Twitter **2021**.
40. TextBlob: Simplified Text Processing. https://textblob.readthedocs.io/en/dev/.
41. Yadav, R.K.; Jiao, L.; Granmo, O.C.; Goodwin, M. Human-level interpretable learning for aspect-based sentiment analysis. The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21). AAAI, 2021.
42. Huang, J.; Meng, Y.; Guo, F.; Ji, H.; Han, J. Weakly-supervised aspect-based sentiment analysis via joint aspect-sentiment topic embedding. *arXiv preprint arXiv:2010.06705* **2020**.
43. 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 preprint arXiv:2103.14804* **2021**.
44. Bose, R.; Aithal, P.; Roy, S. Sentiment Analysis on the Basis of Tweeter Comments of Application of Drugs by Customary Language Toolkit and TextBlob Opinions of Distinct Countries. *International Journal* **2020**, *8*.
45. Sentiment Analysis Tweet dataset.
46. Chamoso, P.; Hernández, G.; González-Briones, A.; García-Peñalvo, F.J. Recommendation of technological profiles to collaborate in software projects using document embeddings. *Neural Computing and Applications* **2020**, pp. 1–8.
47. Dabade, M.S.; others. Sentiment Analysis Of Twitter Data By Using Deep Learning And Machine Learning. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)* **2021**, *12*, 962–970.
48. Gan, C.; Wang, L.; Zhang, Z.; Wang, Z. Sparse attention based separable dilated convolutional neural network for targeted sentiment analysis. *Knowledge-Based Systems* **2020**, *188*, 104827.
49. Haralabopoulos, G.; Anagnostopoulos, I.; McAuley, D. Ensemble deep learning for multilabel binary classification of user-generated content. *Algorithms* **2020**, *13*, 83.
50. Irie, K.; Schlag, I.; Csordás, R.; Schmidhuber, J. Going Beyond Linear Transformers with Recurrent Fast Weight Programmers. *arXiv preprint arXiv:2106.06295* **2021**.
51. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* **2018**.
52. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. Advances in neural information processing systems, 2017, pp. 5998–6008.
53. Hua, Y. Understanding BERT performance in propaganda analysis. *arXiv preprint arXiv:1911.04525* **2019**.
54. Topal, M.O.; Bas, A.; van Heerden, I. Exploring transformers in natural language generation: Gpt, bert, and xlnet. *arXiv preprint arXiv:2102.08036* **2021**.
55. Golestani, M.; Razavi, S.Z.; Borhanifard, Z.; Tahmasebian, F.; Faili, H. Using BERT Encoding and Sentence-Level Language Model for Sentence Ordering. International Conference on Text, Speech, and Dialogue. Springer, 2021, pp. 318–330.

56. Gu, X.; Liu, L.; Yu, H.; Li, J.; Chen, C.; Han, J. On the transformer growth for progressive bert training. *arXiv preprint arXiv:2010.12562* **2020**.
57. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* **2019**.
58. Alzahrani, E.; Jololian, L. How Different Text-preprocessing Techniques Using The BERT Model Affect The Gender Profiling of Authors. *arXiv preprint arXiv:2109.13890* **2021**.
59. González-Carvajal, S.; Garrido-Merchán, E.C. Comparing BERT against traditional machine learning text classification. *arXiv preprint arXiv:2005.13012* **2020**.
60. Jang, H.; Rempel, E.; Roth, D.; Carenini, G.; Janjua, N.Z. Tracking COVID-19 discourse on twitter in North America: Infodemiology study using topic modeling and aspect-based sentiment analysis. *Journal of medical Internet research* **2021**, *23*, e25431.
61. Twitter Rules and policies. https://help.twitter.com/en/rules-and-policies#twitter-rules.
62. Twitter Developer Agreement and Policy. https://developer.twitter.com/en/developer-terms/agreement-and-policy.
63. List of terrorist incidents in 2021. https://en.wikipedia.org/wiki/List_of_terrorist_incidents_in_2021.