electronics

MDPI

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, University of Salamanca, 37007 Salamanca, Spain; guillehg@usal.es (G.H.); chamoso@usal.es (P.C.); corchado@usal.es (J.M.C.)
2 Science Faculty, University of Salamanca, 37008 Salamanca, Spain; Nastaran@usal.es
3 Air Institute, IoT Digital Innovation Hub, Carbajosa de la Sagrada, 37188 Salamanca, Spain
4 Department of Electronics, Information and Communication, Faculty of Engineering, Osaka Institute of Technology, Osaka 535-8585, Japan
5 Pusat Komputeran dan Informatik, Universiti Malaysia Kelantan, Karung Berkunci 36, Pengkaan Chepa, Kota Bharu 16100, Kelantan, Malaysia
* Correspondence: Niloufar.shoeibi@usal.es; Tel.: +34-617-939-365

Abstract: Maintaining a healthy cyber society is a great challenge due to the users' freedom of expression and behavior. This can be solved by monitoring and analyzing the users' behavior and taking proper actions. 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 and filtering component, the interesting profiles are selected for further examinations. Finally, in the contextual analysis component, the contents are analyzed using natural language processing techniques; a binary text classification model (SVM (Support Vector Machine) + TF-IDF (Term Frequency—Inverse Document Frequency) with 88.89% accuracy) is used to detect if a 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 (Feed-Forward Neural Network) with 80% accuracy), because sharing positive opinions about a crime-related topic can threaten society. This platform aims to provide the end-user (the police) with 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, 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

Academic Editor: Amir Mosavi

Received: 30 October 2021
Accepted: 30 November 2021
Published: 10 December 2021

1. Introduction

Huge amounts 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 the interpretation of the data generated by users on social media platforms; it enables interested parties to track the flow of information. It is up to the users to ethically utilize social media platforms; for example, it is possible to share news or learn unethically, such as propagating negative thoughts, violations, racist ideas, etc. A good instance of the misuse of 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 to make 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 others worldwide, and they can share their beliefs, interests, and many broad topics. However, unfortunately, the reach of these networks is extensive, and abusers can bully people according to their race, body image, or religious beliefs [3].

Both victims and nonvictims can suffer from a lack of security, work productivity, loss of money and property, and medical problems. To keep safe, people use extra protection and spend large amounts of money to prevent crime. Moreover, a lack of safety may cause mental and physical suffering and irreparable damages and reduce quality of life. Crime also affects economic wealth, causing victims to fail to be productive at their job. Furthermore, governments have to spend funds on police departments, courts, treatment programs, medical expenses, social workers, and security guards, and a great deal of time for victims, their families, and court trials. Thus, detecting crime can lead to a healthier society and make people's lives better [4].

The first step in the building of a Social Media Analysis (SMA) tool is to capture information from social media platforms. This can be done by using the official APIs (Application Programming Interfaces) [5] of each platform—in the case of Twitter, it has APIs for legally gathering information. All the information related to a tweet is saved in an entity called a tweet object in the JSON (JavaScript Object Notation) format. To extract a significant number of tweets, each tweet object is held in a list and then goes through preprocessing and an analysis of the existing features. This is the first step to start exploring the behavior of the users. Managing the data demands specific policies to be followed [6], which have been considered for this research. After the data are 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 four different components:

1. The data are extracted using the official Twitter API provided by the Twitter developer team. After the data are obtained, the relation of the users is elicited, and a 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 these 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 State of the Art

There are many research studies that have been performed in the area of data analysis and artificial intelligence to detect, optimize, and predict an event in different fields such as anomaly detection [7], profile generation systems 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 behavioral patterns leading to action, from traveling, marketing, and advertising to event detection and crime detection. Below, some of the most recent research works conducted on this topic are reviewed.

Cauteruccio et al. in [9] provided research on the Reddit social media platform, examining three perspectives theoretically and practically. They first explained the dataset that was used in their study; then, they presented the initial results clarifying the subreddit stereotypes. Three macro-categories and some stereotypes for each of them were presented. Finally, three orthogonal taxonomies were employed to assign the discovered stereotypes, and the same process was conducted for the authors' stereotypes. Thus, their platform verified whether Reddit is assertive, and many applications can benefit from subreddit and author stereotypes.

In our previous research, in [10], information extracted from Twitter was used to categorize users. A feature-based study was carried out by combining graph analysis and metadata analysis. It was possible to calculate the importance of nodes, which determined the status of Influencers (the highest importance and related characteristic features, such as 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 research works in the area of detecting malicious accounts, such as spambots and fake followers, have used profile-based and graph-based features; however, in [11], a classification model was built which only considered the account's tweets. As a result, the highest accuracy was obtained using TF-IDF features and the XGBoost algorithm, at 95.55%. In comparison, Word2vec features and the XGBoost algorithm achieved an accuracy of 95.2% in malicious vs. genuine account detection.

N. Shoeibi et al. in [12] proposed a system to detect cyber victimization. They aimed to catch certain types of behavior that indicate suicide, self-harm, and cyberbullying and created a dataset using the official news published about these events. The results showed 96% accuracy in the SVC tweet classification model.

M. Hasan et al. in [13] streamed tweets to gather 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 [14] identified Twitter messages with the potential of promoting illegal services and used by minors by applying natural language processing techniques. 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 are 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 results in a higher performance than CNN.

The framework proposed by Zaheer Abbass et al. [15] consists of three stages: data preprocessing, a classifying model builder, and prediction. As the prediction models, Multinomial Naïve Bayes (MNB), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM)wereused. First, the model classifies the data into different categories of crime. Later, an N-gram language model is used with machine learning algorithms to discover the n 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, with recall and F-measure scores of 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% [16].

In [17], a case study was carried out in India, where Twitter data were 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, a Twitter part-of-speech tagger, a Markov Model of first-order entropy, was used for parts-of-speech in online conversational text. Then, Brown Clustering was utilized for a large set of unlabeled tweets. According to different locations, the results were 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 [18], 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 were selected between affine and hierarchical clustering; Latent Dirichlet Allocation captured topics in a pipeline of Doc2Vec and K-means and used a multi-step polarity index classification and its time series analysis. First, the authors 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 [19], studied the theoretical, practical, and methodological implications of online hate speech and the sentiments of tweets and discussed the manual and computational techniques required to investigate the stream of Twitter messages in Spanish, with 24,254 samples before and after the declaration of the Spanish government 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 sentimental viewpoint of the tweets became more positive. In their model, they used topic modeling and sentiment analysis for the Spanish language.

In [20], 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 were compared by correlating the time-series features extracted from each timeline using Dynamic Time Warping. The audience was the followers and the users who interacted with the main user's content. A higher overlap between the sets of audiences represented a greater similarity of the users. The text of tweets was also compared between two profiles, and the number of tweets that were the same was collected; the content similarity was calculated using TF-IDF and Cosine Similarity.

H. Yin et al., in [21], investigated how to learn to represent brief texts for text clustering. They examined the available pre-trained models such as 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 accuracy by 14% when clustering brief text.

In the next section, the proposed method is explained in detail. AI-Crime Hunter investigates the crime on Twitter from three aspects; the network of interactions, behavioral patterns, and contextual analysis. Hence, it decides if a user has criminal ideologies and behavior based on the knowledge derived from these three aspects. Therefore, Table 1 indicates that the solutions proposed in each related article, as explained this section, cover

each type of aspect. As can be observed, AI-Crime Hunter covers a broader point of view for detecting the expression and propagation of crime.

Table 1. The summary of the state of the art and the proposed solutions for each problem.

Article	Network of Interactions	Behavioral Patterns	Contextual Analysis
Cauteruccio et al. [9]	X	X	-
N. Shoeibi et al. [10]	X	X	-
F. N. Pakaya et al. [11]	X	-	X
N. Shoeibi et al. [12]	-	-	X
M. Hasan et al. [13]	-	-	X
S. L. Granizo et al. In [14]	-	X	-
Zaheer Abbass et al. [15]	-	X	X
Sangeeta Lal et al. [16]	-	-	X
T. Vo et al. [17]	-	X	X
S. Mendon et al. [18]	X	-	X
C. Arcila-Calderón et al. [19]	-	-	X
N. Shoeibi et al. [20]	-	X	X
H. Yin et al. [21]	-	-	X
AI-Crime Hunter (Proposed System)	X	X	X

3. AI-Crime Hunter 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 [22].

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 us to solve the research challenges;

(Q1) How can the agreement level of a user with criminal ideologies be calculated?

(Q2) How can the connections between users be defined and influential users found?

(Q3) Which attributes determine the behavioral consistency of a profile?

(Q4) How can a criminal event be detected?

Answering all the above questions is the critical point in the development of a platform to measure the popularity of illegal content. The architecture of the proposed platform is presented in Figure 1. In the AI-Crime Hunter platform, firstly, the data are extracted from Twitter using official Twitter APIs. It is crucial to follow Twitter's policies regarding data publication. It is necessary to anonymize the information about the profiles; however, the end-user of this platform is the police force, and profile suspension only happens in very high-risk cases. After the first step, the data are analyzed and converted into meaningful information. This architecture consists of five different components, which 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 conducted, 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 and filtering**: Nine behavioral categories are 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' timelines 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. After applying the binary text classification model, the topic of each tweet is categorized into two classes: crime-related or not. This classification model is trained on an available labeled dataset, called the Global Terrorism Dataset (GTD) [23], which is 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, are applied on the text of tweets, trained on an available labeled dataset of the tweets and their sentiments [24]. However, the results show 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 labels 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.

This section has been divided into different subsections to explain each component more precisely. Section 3.1 explains the data extraction methods. Section 3.2 takes a deeper look at graph network analysis and explains the details of the design. Then, in Section 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 Section 3.4, nine filters are applied to the obtained information from the previous sections, and the data are reduced. Finally, in Section 3.5, the text of each tweet on a user's timeline is subjected to 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. Then, in the metadata analysis component, the recent 3200 tweets posted on each profile's timeline are extracted and delivered to extract the advanced features.

The first step of using 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 consider [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 reveal a great deal. Analyzing the interactions between users opens the door to recognizing the behavioral patterns of the users [25].

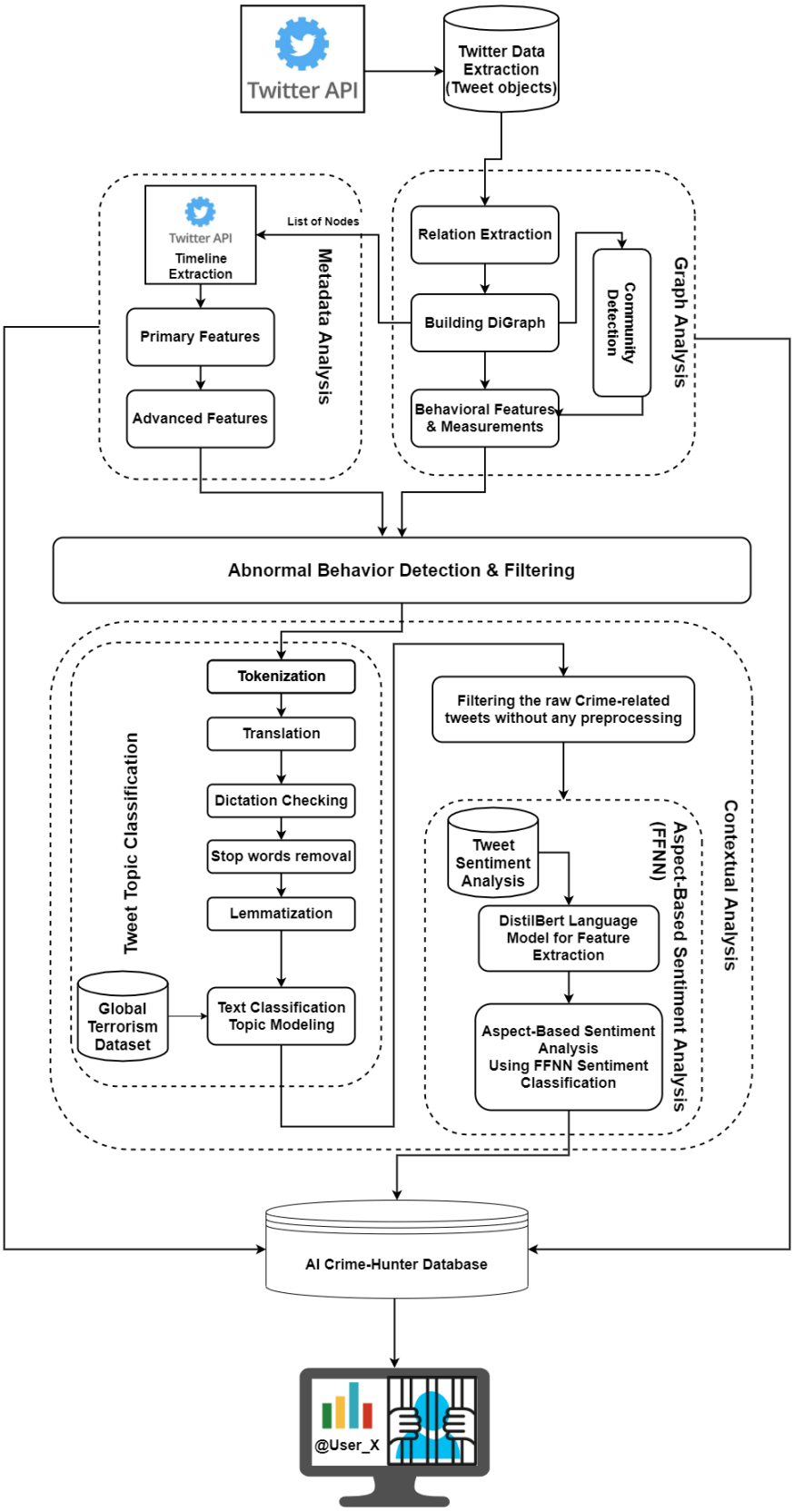

Figure 1. The architecture of the AI-Crime Hunter platform.

3.2.1. Graph Structure and Measurements

Twitter has provided some content-sharing strategies for users to interact with each other by employing these strategies efficiently. When analyzing the intercommunication of

the users on social media platforms, graph network analysis is a great solution to solve this problem. When building a graph network, the nodes and edges need to be defined. Users (The screen-names and IDs of users were chosen because these represent unique values that hold identical information of the users) are considered nodes of a graph, and retweeting, quoting, replying, and mentioning are considered to define the users' relationships 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, these data are used to create the graph network of interactions. Analyzing the graph helps to create new attributes [26] that mostly show the importance level of the nodes—i.e., users—in the network, as represented in Table 2.

The implementation of the directed graph has been conducted with Python, using the NetworkX library [27]. It provides functions for estimating structural and centrality measurement.

Table 2. Features extracted from graph network analysis.

Attributes	Definition
Eccentricity	The maximum shortest distance of one node from others. The lower the eccentricity, the greater 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 simply 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, independent 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 3 represents the notation used in the formulations presented in Table 2.

Table 3. The notation used in the formulations of Table 2.

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
δ_{st}	Amount of quickest routes between (s, t)
$\delta_{st}(u)$	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 with a similar behavior or characteristic shape a community, the members of a tennis club, the students of a programming class, or people above the age of 50. The members of a community have at least one thing in common; however, it is sometimes 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 Section 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, which considers a graph's structure. Thus, based on 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 [28].

The **Girvan–Newman** method is an algorithm that is used to perform 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 a scalability and computational complexity of $(O(m^3))$ for weighted and $O((m^3) + (m^3)logm)$ for unweighted graphs [29]. Algorithm 1 was implemented in Python by utilizing the NetworkX library [30].

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 basic information to be extracted 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 the text, date of creation, number of favorites, number of retweets, etc., as well as information on the user who is posting a tweet, such as the number of followers, number of users followed, date of creation of the account, the total number of tweets, lists that the user belongs to, etc.

3.3.1. Primary Features

Once the preliminary information from tweet objects has been extracted, as presented in Tables 4 and 5, it is given to the next level: advanced features extraction [31].

Table 4. Primary features extracted from the Twitter Data Dictionary related to a user's 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 5. 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 the date time as the time axis, the attributes of the users can be transferred into time-series values. These values show the consistency of the behavior, the users' activity level, and the steadiness of user engagement. Table 6 explains the advanced features that are generated in this step.

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

Table 6. 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 that have been used
URL	Mean number of tweets that includes a URL

3.4. Abnormal Behavior Detection and 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 to remove 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; based on all the variables obtained, a series of heuristic rules are defined. These categories are listed below.

- ***Old spreader:*** The type of profile that publishes a large number of tweets on a given day, 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 retweeted a large number of tweets on a recent and specific day while barely 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 large number of daily tweets.
- ***Influencers (III):*** An influencer class with many favorites and retweets that also publishes constantly, with high centrality measures.
- ***Bots:*** A type of profile that posts several tweets each day and the behavioral ratios derived from its behavior is not like humans.

In the *abnormal behavior detection and filtering* component, the users with suspicious behavior are selected to be studied in more detail by giving them as the input to the next component, 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. To do this, each tweet on the user's timeline goes through different steps in two sub-components.

3.5.1. Topic Classification

Preprocessing was conducted in Python using NLTK [32], Textblob [33], re [34], etc. Preprocessing consists of translation, tokenization, dictation checking, removing stop words, and lemmatization.

Tokenization is the technique used 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 to understand different languages. In this research, the Google Translate API [35] has been used because it contains a vast range of other languages; it is also easy and cheap to use. Table 7 represents a comparison between some popular translation APIs.

Table 7. The comparison between translation APIs.

API	Languages	Pricing	Popularity	Latency
Google Translate API [35]	108	20$	9.9/10	493 ms
IBM Watson Language Translator API [36]	39	20$	8.1/10	256 ms
Yandex Translate API [37]	93	6$	0.2/10	127,262 ms

Google Translate API has the highest user satisfaction, and it is easiest to implement and use; therefore, this API was chosen for the translation. Dictation checking is mandatory as, due to the character limitation of the tweets, users tend to shorten 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.); finally, lemmatization is the process of restoring different versions of a word into their root. 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 determine whether the user agrees or disagrees with the crime by applying sentiment analysis on the selected contents.

A text classification model was implemented and trained on the Global Terrorism Dataset (GTD) [23] to apply topic modeling. This dataset is an open-source dataset containing information on 180,000 terrorist attacks worldwide from 1970 until 2017. Figure 2 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.

Globally, all regions have seen an increase in the average impact of terrorism in recent years compared to the start of the 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 labels. This enables us to train a text classification model and, more importantly, evaluate it by comparing the actual values and the predicted labels. For the deployment of a text classification model, the text feature extraction needs to be performed before all other steps.

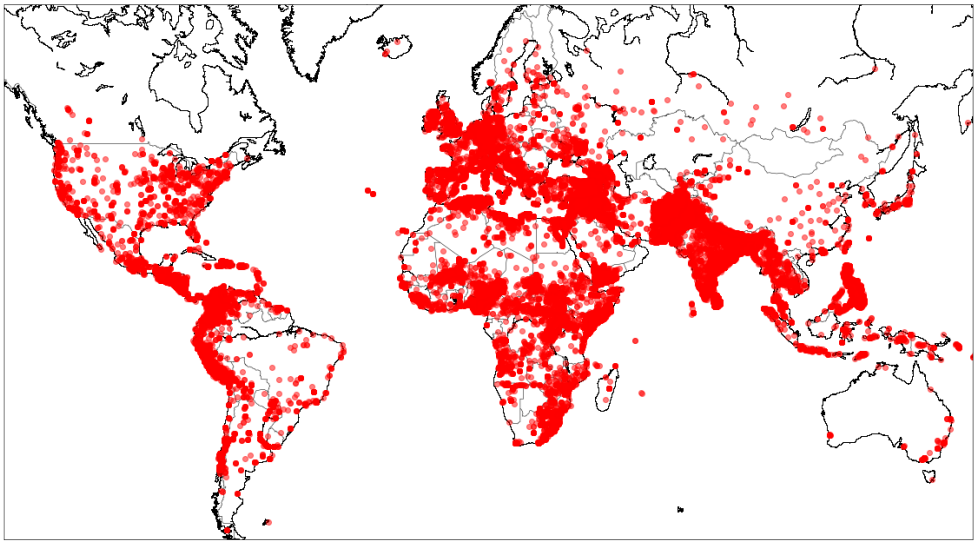

Figure 2. The distribution of the terrorist attacks all over the world.

Figure 3 demonstrates the process of the text topic classification. 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 are applied; for this, six different classification models were selected.

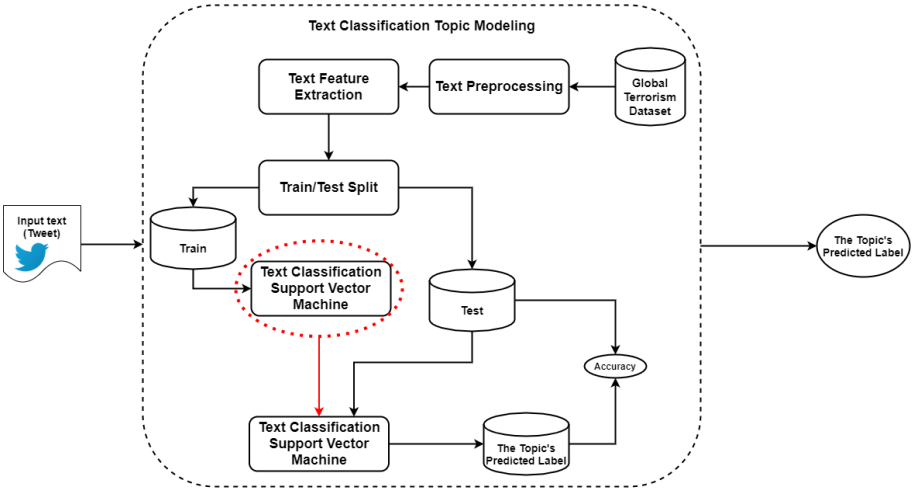

Figure 3. The text classification topic modeling.

Vectorization is a method to transfer text into numerical data, as required to apply any machine learning algorithm. The count vectorizer uses the number of the times that a 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 [38].

In contrast, the Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer [39] is a method that considers the significance of a word by calculating the term frequency, showing how frequently a word appears in the document. IDF, which is the weight of rare words, refers to words that rarely appear in the document. By multiplying these two values, the importance of the words is obtained. The equation below represents the TF-IDF formula.

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

where t is a term (word), d is the document (set of words), $df(t)$ is the document's frequency of t, N is the count of the corpus, and $corpus$ is the total document set.

To find the best result, two strategies of text feature extraction—the count vectorizer and the TF-IDF method—were applied, and the results are compared with each other. Table 8 represents the results of the feature extraction and text classification models' accuracy.

Table 8. 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% accuracy. Therefore, after running the text classification models and detecting the topics of each tweet, the tweets related to crime were filtered and handed over to the next step—the aspect-based sentiment analysis—to discover the polarity and subjectivity of the users 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 perform sentiment analysis [40], including three main solutions: rule-based [41], feature-based [42], and embedding-based methods [43,44]. The aspect-based sentiment analysis model is implemented using rule-based models to measure an input text's subjectivity, indicating if a tweet is a fact or an opinion, and transformer-based sentiment analysis to understand if a tweet induces positive or negative emotions.

To measure the subjectivity of the tweets, the Textblob algorithm [45], which is implemented as a library called TextBlob [41], 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. The polarity is a number between -1 and $+1$, which shows the text's positivity or negativity; i.e., the closer the text is to -1, the more negative, a value closer to $+1$ shows greater positivity, and 0 represents neutral. However, the subjectivity shows if the input text is closer to an opinion or a fact. Subjectivity, as utilized in this proposed model, is between 0 and 1, where 0 designates facts and 1 indicates opinions; the closer the value to 1, the more likely the text is to be an opinion and vice versa. In addition, 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 made possible. As the goal of AI-Crime Hunter is to analyze and measure the agreement level to criminal topics, the tweets are passed through the topic text classification level, which is trained on the GTD Dataset to classify the input text (tweets) into two categories: crime-related and

non-criminal. Consequently, the sentiment analysis is only performed on the crime-related tweets by filtering these tweets.

The sentiment analysis model was trained with a dataset of 1.6 million tweets, which is an open-source dataset [46] gathered by the CS224N project of Stanford [24]. It is a balanced dataset, with the tweets and their related labels indicating the polarity of each record.

Two different strategies for feature extraction were applied to the dataset, and the extracted features were used to train 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 performed, and the dataset is subjected to tokenization, removing stop words, lemmatization, etc. Then, a Word2vec model is trained using the vocabularies existing in this dataset to save the syntactical information of the words by turning them into vectors in order not to lose the concept of the words and secret relationships between the words [47]. Table 9 represents the similar words derived using the Word2vec model.

Table 9. 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 sets, and this way of using the word embedding model is considered as a method of text feature extraction.

After building the Word2vec model [48], the embedding layer [49] was created using the embedding matrix, using the words and the values of the vectors derived from the Word2vec model. The embedding matrix was defined as the weights of an embedding layer. Then, a Long-Term Short Memory (LSTM) network [50] combined with dense layers and dropouts was trained with the training datasets. The accuracy and the loss of the model in the 15 epochs of training are presented in Figure 4.

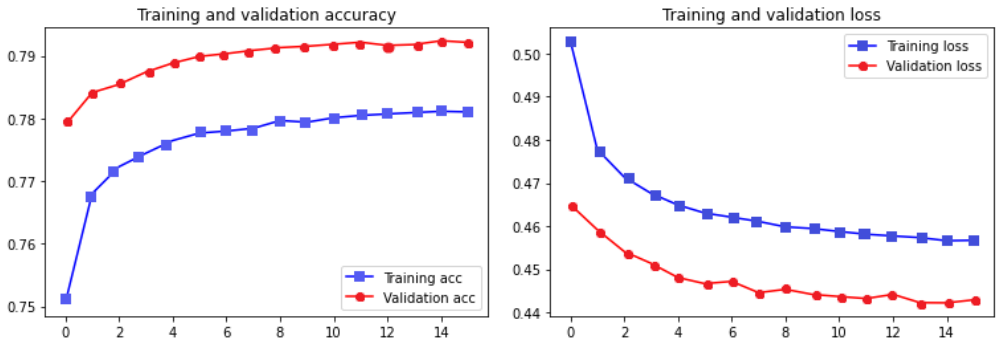

Figure 4. Training and validation accuracy and loss.

Table 10 represents more information about the quality of the classification.

Table 10. The results of the aspect-based sentiment analysis with Word2vec and LSTM.

	Precision	Recall	f1-Score	Support
Negative	0.80	0.78	0.79	160,000
Positive	0.79	0.80	0.80	160,000
Accuracy			0.79	320,000
Macro avg	0.79	0.79	0.79	320,000
Weighted avg	0.79	0.79	0.79	320,000

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

The big drawback o Recurrent Neural Networks and LSTMs are that the data need to go through the networks 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 these networks handle the text more effectively. Liu et al. propose a solution to the vanishing gradient problem in [51]. However, both networks are very complex and need a long time for training and to become effective [52].

Transformers work with a different mechanism. They are attention-based models, which means that they are able to work in parallel, so the text does not need to be input word by word sequentially—they can pass all the sentences at the same time [53]. This ability allows them to perform much faster than RNNs, and they are deeply bidirectional. Figure 5 shows a mechanism of the transformers. In [54], all the details of the architecture of transformers are presented. However, from 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, encapsulating the meanings of the words, meaning that similar words have closer vector 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 [55]. 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 [56].

In this study, the encoding part has been used for text feature extraction in order to turn the sentence into its respective vector [57].

Figure 6 represents the architecture of the encoding part of the transformers. This can solve many different problems requiring an understanding of the language, such as 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 [58]. 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.

In this work, BERT is utilized for feature extraction. Moreover, a lighter, faster, and cheaper version of BERT, which is called DistilBERT [59], is used. DistilBERT is 40% of the size of standard BERT, saving 97% of its language understanding capacity, but is 60% faster. The output vector size is 768, meaning that each sentence will have a fixed-size vector with 768 values.

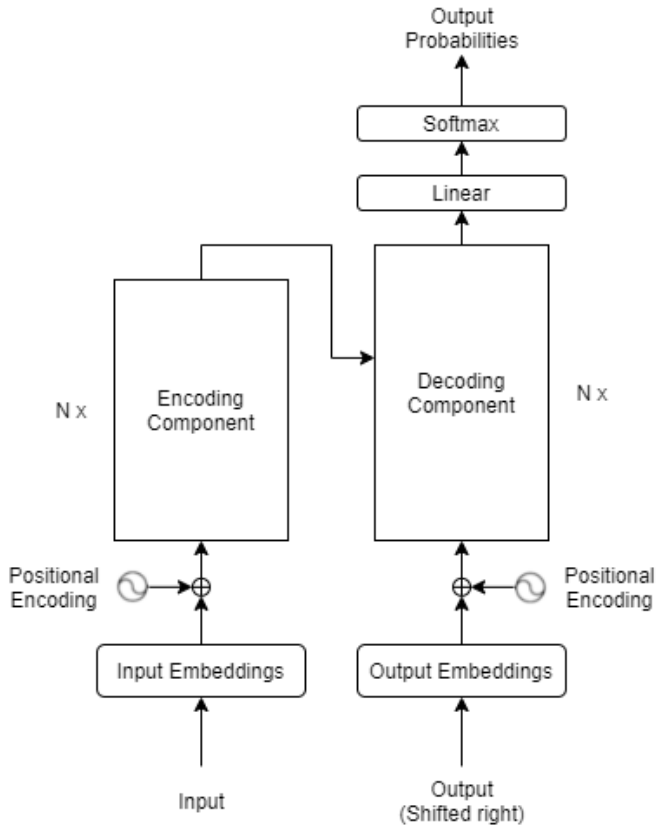

Figure 5. The architecture of the transformers.

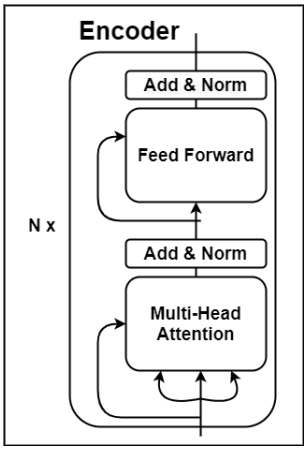

Figure 6. The architecture of the encoder part of the transformers.

It is essential to mention that the text preprocessing for BERT models is different from the traditional text preprocessing in terms of its details and implementations [60,61].

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 that it can vectorize 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 benefit against the LSTM model.

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

Furthermore, Table 11 represents the results of the aspect-based sentiment analysis using the DistilBERT method and Feed-Forward Neural Network.

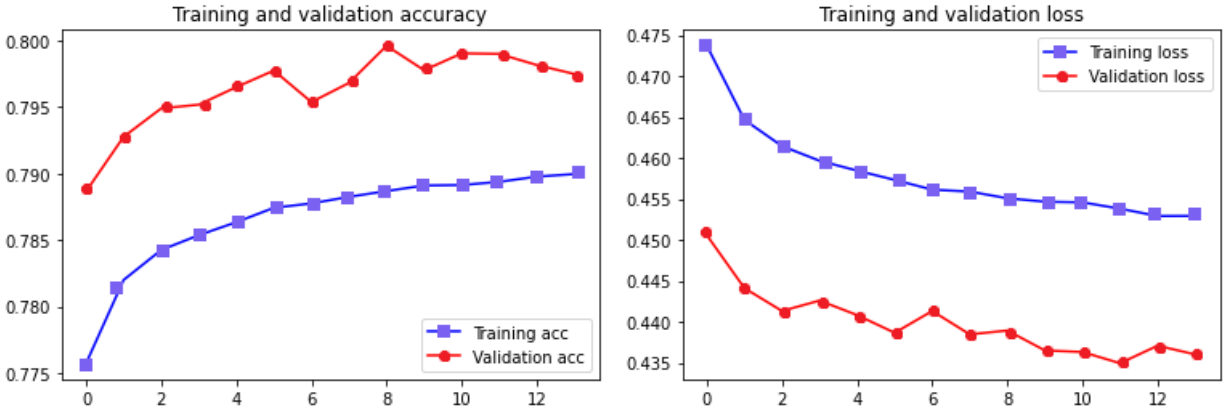

Figure 7. Training and validation accuracy and Loss of DistilBERT + FFNN.

Table 11. The results of the aspect-based sentiment analysis with the DistilBERT method and Feed-Forward Neural Network.

	Precision	Recall	f1-Score	Support
Negative	0.80	0.79	0.80	160,000
Positive	0.79	0.80	0.80	160,000
Accuracy			0.80	320,000
Macro avg	0.80	0.80	0.80	320,000
Weighted avg	0.80	0.80	0.80	320,000

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, and, on the other hand, the simplicity of Feed-Forward Neural Networks makes the second approach a better choice; thus, it was selected for further analysis.

To present some of the outputs of the contextual analysis component of the system, Table 12 presents some extracted tweets and the values of the corresponding results of the sentiment analysis algorithms. As 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 with illegal content.

Table 12. 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 heroes!"	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! People 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 to interpret a sentence's sentiment enables us to understand the emotion of the input text more precisely; this procedure is aspect-based sentiment analysis. It is usually used when sentiment analysis is performed for a specific subject [62]; for example, when a company wants to understand what customers think about its products. Alternatively, in general, ion this research, we are interested in what people think about crime-related topics.

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

4. Results and Case-Study

Twitter's purpose is to serve public conversation. It is significant for its developers to understand their rights and thoroughly know how much of their information is available to others. On the other hand, there are restricted policies [63] regarding Twitter data for researchers and product holders who use Twitter data and analyze them. Due to Twitter's data publishing policies, promulgating any private information needs the user's permission directly. This includes physical location information, identity details, contact details, financial account information, other personal data, biometric data, and medical records.

Analyzing Twitter data may not be as challenging as validating it due to Twitter's highly restricted rules and policies, which are called the Developer Agreement [6]. It is not permitted to save and retrieve the information without the consent of Twitter users, which makes it very challenging to build a benchmark or publish a dataset for further studies and evaluate the general performance of the proposed model. However, as with any scientific article, it is mandatory to assess the proposed method. Due to the lack of available benchmarks on Twitter, it is very difficult to evaluate the performance of the whole system; therefore, each component of the system has been examined separately, and also the first author has created a Twitter profile and consented to publish their data. Thus, to some extent, the results of the proposed technique can be seen to be valid. 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 permitted to reveal the true identity of the users. Therefore, in the results part, instead of using the screen-name of the users, identities are anonymized, such as @User_X. The result of the platform is a list of profiles that are suggested for suspension to stop crime propagation.

According to the last update of Wikipedia's most recent terrorist incidents in 2021 [64], on 4 and 5 June, 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 4 and 5 June 2021 shows the results. A search with the terrorism-related keywords to this attack batched the tweets associated with this attack. Therefore, 2000 tweets with the keywords of "terrorist attack in Burkina Faso" and "mass shooting in Burkina Faso" were extracted. The proposed method aimed to detect the profiles expressing a high level of agreement with illegal content. First, the graph analysis component extracted the interactions of people with each other and created a list of the unique users, called a node list. After extracting data, the intercommunication of the users with each other was detected, and a respective graph network is built. Table 13 shows a sample of the interconnections between users. In this table, the screen-name of the users represents the nodes, and the relationships, retweets, quotes, mentions, and replies define 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.

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

Table 13. The interconnection of users, with 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

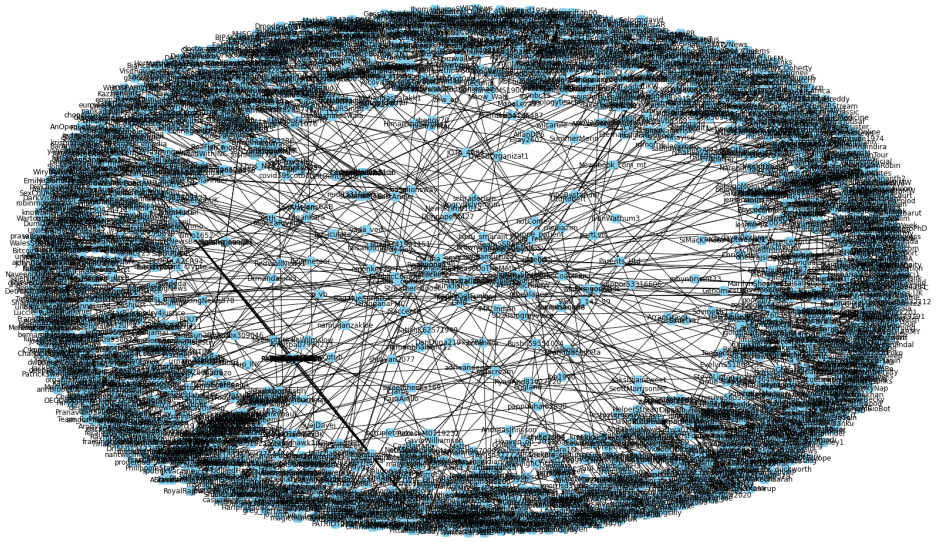

Figure 8. A graph sample of the users tweeting about the mass shooting in Burkina Faso.

Then, the graph features are measured, and the community detection algorithm is run. Therefore, an overview of the nodes and the communities is achieved. In parallel, when the node list is ready, it is handed to the next component: the metadata analysis component, which consists of timeline extraction, as well as primary and advanced feature extractions. By considering all the information collected from graph network analysis and the metadata analysis, the wide searching area is narrowed by applying the filters to detect users with behaviors of interest. From the 2000 tweets collected from the query, 1825 unique users were found; after using the filters, the searching area was narrowed into 543 profiles, which is a reduction to 27%.

In the next step, the timeline of these candidate profiles was passed through the process of contextual analysis to find the profiles spreading crime-related content and calculate their level of agreement with the illegal content. To show how the platform works, step by step, the information of the process is explained. Nevertheless, because of the restricted Twitter policies, which do not allow data and identities to be published without consent from the users themselves, a new Twitter profile was created to show a successful case, and some crime-related and non-criminal content was 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, was performed using natural language processing (NLP) techniques, discussed in more detail in the previous section. First, each tweet was passed through preprocessing, translation to English, tokenization, dictation checking, and lemmatization; then, a text classification model using SVM was

used. Each tweet was labeled as crime-related or non-criminal. Next, the crime-related tweets were separated and passed through the process of sentiment analysis. Aspect-based sentiment analysis is used to inform the end-user—police organizations—to what extent users agree or disagree with the illegal content according to two attributes; polarity and subjectivity. Polarity shows the positivity level, and subjectivity represents if a text is more of a fact or an opinion. A positive opinion about a crime-related topic represents a high level of agreement with illegal activities. A Twitter profile showing a high agreement level with criminal issues is a threat to society.

In this case study, contextual analysis was applied to the published profile. Table 14 represents the results of the contextual analysis component involved in some of the tweets.

Table 14. 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 project 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 is shown.

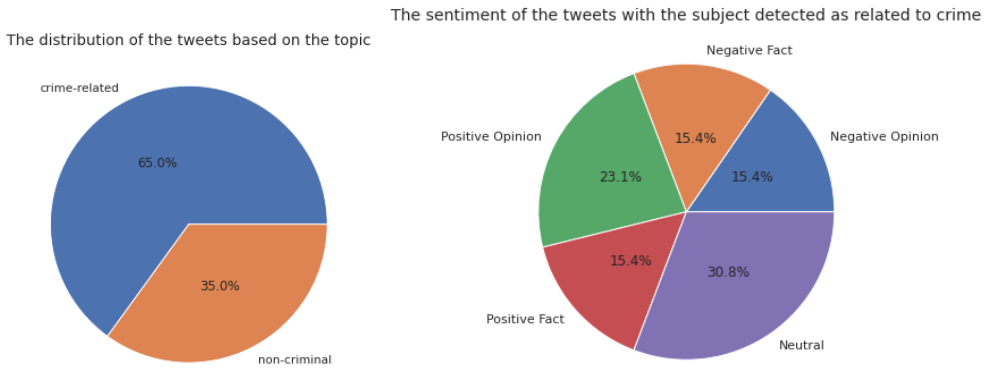

Figure 9. The pie chart representing the percentage of the crime-related and non-criminal tweets and the measured sentiment of the crime related tweets.

As explained before, a positive opinion about illegal content exposes a high level of agreement with the crime. The results of the contextual analysis proves 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 crime-related topics, which shows a high level of agreement with crime.

In the next section, the conclusions are presented; besides, the goal of implementing the AI Crime-Hunter is explained. In addition, plans for improving the system are proposed.

5. Conclusions 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 criminal subjects and make suggestions for account suspension. The final goal is to stop spreading criminally positive opinions to reduce the crime rate eventually. Many different AI techniques have been employed in this platform, created by a mixture of AI experts to detect crime on Twitter. Graph network analysis is used to map the intercommunications of the users with each other and calculate various graph measurements to find the most and least essential nodes—those that are centric and are involved in the shortest paths of one node to all others. Each user's timeline was thus extracted. By analyzing these data, primary features and more advanced features such as the time-series-related attributes were calculated, showing the behavioral consistency of the profile's activity during the time. Then, by summing up all the previously gathered information, nine behavioral filters were designed to narrow the data size for the input of the next step.

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

After the preprocessing, topic modeling was performed with an SVM text classification model trained with the Global Terrorism Dataset, a public dataset [23], and the TF-IDF vectorizer for text feature extraction. This model was able to detect if a tweet was related to illegal content with an accuracy of 88.89%. After distinguishing the legal and illicit content, this information helped us to observe how many tweets on the timeline were related to this category, which is helpful.

Then, aspect-based sentiment analysis was performed to measure the profile's agreement level to the illegal content. The mechanism consisted of DistilBERT as text feature extraction method, which transferred each tweet into a fixed-sized vector of 768; then, each vector with its respective label was passed through a Feed-Forward Neural Network. This model was trained on the dataset of 1.6 million labeled tweets, from an open-source dataset collected by Stanford [46]. Then, it was used to predict the sentiment of the extracted tweets. The proposed sentiment classification performed with 80% accuracy. Besides, for measuring the subjectivity of tweets, the Textblob algorithm was utilized. This is a rule-based sentiment analysis algorithm that calculates subjectivity, showing if a text 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, which is a unique idea to crime.

Mathematically modeling the research questions is too complicated because the answer to each question is defined as an algorithmic procedure, which is described deeply in each subsection. However, the solution for each question is addressed below:

(A1) In the ***contextual analysis component***, a user's agreement level is measured by the process of aspect-based sentiment analysis, meaning that each tweet posted on a user's timeline goes through a text classification model that predicts if the tweet is crime-related or not. Then, the sentiment of each crime-related tweet is measured. Finally, the number of the "***positive opinions***" in crime-related posts indicates the agreement level of the user to the crime.

(A2) In the ***graph analysis component,*** the interconnections between the users are studied by considering the users as nodes and the communication strategies (retweets, replies, quotes, etc.) as edges of the directed graph. Then, the most influential nodes are selected for further examinations.

(A3) In the ***metadata analysis component***, many primary and advanced features defining the historical behavior habits of the users are calculated, which are explained in detail in the respective subsection.

(A4) Crime detection and prediction are very challenging. This involves a stochastic chain of events, actions, and features that affect each other. Utilizing the whole proposed system helps to find users who share positive opinions about criminal topics. An incredible amount of knowledge is derived by highlighting detected users in the network of all users in a query to monitor the community they belong to and the audience of suspicious users. For example, a suspicious node detected by the AI-Crime hunter, whose recent crime-related tweet has a significant level of interactions (replies) with a positive sentiment, reveals a profile promoting crime.

This platform aims to provide an overview of users' behavioral information and activity patterns and the users' agreement level with illegal subjects. A human expert is required to make a decision about the suspension of a profile. In addition, by detecting the communities on the graph network, the immoral content sharing of users can be monitored to stop the further expansion of depravity.

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

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

Funding: This research received no external funding.

Data Availability Statement: Due to the restricted policies of data publication for Twitter, the data are confidential; the steps of data extraction from this platform using the official Twitter API have been clearly explained, but we do not have 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 at "https://www.kaggle.com/START-UMD/gtd" (accessed on 1 November 2021). Furthermore, the aspect-based sentiment analysis model is on a public dataset, gathered and labeled by Stanford and available at "https://www.kaggle.com/kazanova/sentiment140" (accessed on 1 November 2021).

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 the 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.

References

1. Jiang, P.; Van Fan, Y.; Klemeš, J.J. Data analytics of social media publicity to enhance household waste management. *Resour. Conserv. Recycl.* **2021**, *164*, 105146. [CrossRef]
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: Berlin/Heidelberg, Germany, 2021; pp. 149–159.
3. Simović, M.; Kuprešanin, J. Criminal Offenses with Elements of Violence-Psychology of Crimw and Abuse of Power. *Knowl. Int. J.* **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 Soc.* **2020**, *48*, 357–388. [CrossRef]
5. Rate Limits | Docs | Twitter Developer. Available online: https://developer.twitter.com/en/docs/twitter-api/v1/rate-limits (accessed on 1 November 2021).
6. Twitter Agreement and Policy | Twitter Developer. Available online: https://developer.twitter.com/en/developer-terms/agreement-and-policy (accessed on 1 November 2021).
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**, *452*, 498–507. [CrossRef]
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. *J. Ambient Intell. Humaniz. Comput.* **2020**, *11*, 4583–4592. [CrossRef]
9. Cauteruccio, F.; Corradini, E.; Terracina, G.; Ursino, D.; Virgili, L. Investigating Reddit to detect subreddit and author stereotypes and to evaluate author assortativity. *J. Inf. Sci.* **2020**. [CrossRef]
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. In Proceedings of the International Symposium on Distributed Computing and Artificial Intelligence, L'Aquila, Italy, 16–19 June 2020; Springer: Berlin/Heidelberg, Germany, 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. In Proceedings of the 2019 Fourth International Conference on Informatics and Computing (ICIC), Semarang, Indonesia, 16–17 October 2019; pp. 1–5.
12. Shoeibi, N.; Shoeibi, N.; Julian, V.; Ossowski, S.; Arrieta, A.G.; Chamoso, P. Smart Cyber Victimization Discovery on Twitter. In Proceedings of the Sustainable Smart Cities and Territories International Conference, Doha, Qatar, 27–29 April 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 289–299.
13. Hasan, M.; Orgun, M.A.; Schwitter, R. Real-time event detection from the Twitter data stream using the TwitterNews+ Framework. *Inf. Process. Manag.* **2019**, *56*, 1146–1165. [CrossRef]
14. 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. [CrossRef]
15. Abbass, Z.; Ali, Z.; Ali, M.; Akbar, B.; Saleem, A. A Framework to Predict Social Crime through Twitter Tweets by Using Machine Learning. In Proceedings of the 2020 IEEE 14th International Conference on Semantic Computing (ICSC), San Diego, CA, USA, 3–5 February 2020; pp. 363–368.
16. 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. [CrossRef]
17. 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. *J. Intell. Fuzzy Syst.* **2020**, *38*, 4287–4299. [CrossRef]
18. 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. *Inf. Syst. Front.* **2021**, *23*, 1145–1168. [CrossRef]
19. 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. [CrossRef]
20. Shoeibi, N.; Shoeibi, N.; Chamoso, P.; Alizadehsani, Z.; Corchado, J.M. Similarity Approximation of Twitter Profiles. *Preprints* **2021**. [CrossRef]
21. Yin, H.; Song, X.; Yang, S.; Huang, G.; Li, J. Representation Learning for Short Text Clustering. *arXiv* **2021**, arXiv:2109.09894.
22. Bahar, H.M. Social Fmedia and disinformation in war propaganda: How Afghan government and the Taliban use Twitter. *Media Asia* **2020**, *47*, 34–46. [CrossRef]
23. Global Terrorism Database. Available online: https://www.kaggle.com/START-UMD/gtd (accessed on 1 November 2021).
24. Go, A.; Bhayani, R.; Huang, L. *Twitter Sentiment Classification Using Distant Supervision*; CS224N Project Report; Stanford University: Stanford, CA, USA, 2009; Volume 1, pp. 1–6.
25. Sujon, M.; Dai, F. Social Media Mining for Understanding Traffic Safety Culture in Washington State Using Twitter Data. *J. Comput. Civ. Eng.* **2021**, *35*, 04020059. [CrossRef]
26. Hasson, S.T.; Hussein, Z. Correlation among network centrality metrics in complex networks. In Proceedings of the 2020 6th International Engineering Conference "Sustainable Technology and Development" (IEC), Erbil, Iraq, 26–27 February 2020; pp. 54–58.
27. Hagberg, A.; Swart, P.; Chult, D.S. *Exploring Network Structure, Dynamics, and Function Using NetworkX*; Technical Report; Los Alamos National Lab. (LANL): Los Alamos, NM, USA, 2008.

28. Li, S.; Jiang, L.; Wu, X.; Han, W.; Zhao, D.; Wang, Z. A weighted network community detection algorithm based on deep learning. *Appl. Math. Comput.* **2021**, *401*, 126012. [CrossRef]
29. Arasteh, M.; Alizadeh, S. A fast divisive community detection algorithm based on edge degree betweenness centrality. *Appl. Intell.* **2019**, *49*, 689–702. [CrossRef]
30. Girvan-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).
31. Tweet Object. Available online: https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet (accessed on 1 November 2021).
32. Loper, E.; Bird, S. Nltk: The natural language toolkit. *arXiv* **2002**, arXiv:0205028.
33. Loria, S. *Textblob Documentation*; Release 0.15; 2018. Available online: https://buildmedia.readthedocs.org/media/pdf/textblob/latest/textblob.pdf (accessed on 1 November 2021)
34. Chapman, C.; Stolee, K.T. Exploring regular expression usage and context in Python. In Proceedings of the 25th International Symposium on Software Testing and Analysis, Saarbrücken, Germany, 18–20 July 2016; pp. 282–293.
35. Google Translate API. Available online: https://cloud.google.com/translate (accessed on 1 November 2021).
36. IBM Watson Language Translator API. Available online: https://cloud.ibm.com/apidocs/language-translator (accessed on 1 November 2021).
37. Yandex Translate API. Available online: https://yandex.com/dev/translate/ (accessed on 1 November 2021).
38. Indarapu, S.R.K.; Komalla, J.; Inugala, D.R.; Kota, G.R.; Sanam, A. Comparative analysis of machine learning algorithms to detect fake news. In Proceedings of the 2021 3rd International Conference on Signal Processing and Communication (ICPSC), Coimbatore, India, 13–14 May 2021; pp. 591–594.
39. Rawat, M.S.; Srivastava, A.; Aggarwal, S. Detection of Fake News Using Machine Learning. *Int. J. Eng. Appl. Phys.* **2021**, *1*, 205–209.
40. 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. [CrossRef]
41. TextBlob: Simplified Text Processing. Available online: https://textblob.readthedocs.io/en/dev/ (accessed on 1 November 2021).
42. 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.
43. Huang, J.; Meng, Y.; Guo, F.; Ji, H.; Han, J. Weakly-supervised aspect-based sentiment analysis via joint aspect-sentiment topic embedding. *arXiv* **2020**, arXiv:2010.06705.
44. 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**, arXiv:2103.14804.
45. 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. *Int. J.* **2020**, *8*. [CrossRef]
46. Sentiment Analysis Tweet Dataset. Available online: https://www.kaggle.com/kazanova/sentiment140 (accessed on 1 November 2021).
47. Chamoso, P.; Hernández, G.; González-Briones, A.; García-Peñalvo, F.J. Recommendation of technological profiles to collaborate in software projects using document embeddings. *Neural. Comput. Appl.* **2020**, 1–8. [CrossRef]
48. Dabade, M.S.; Sale, M.D.; Dhokate, D.D.; Kambare, S.M. Sentiment Analysis of Twitter Data by Using Deep Learning and Machine Learning. *Turk. J. Comput. Math. Educ. (TURCOMAT)* **2021**, *12*, 962–970.
49. Gan, C.; Wang, L.; Zhang, Z.; Wang, Z. Sparse attention based separable dilated convolutional neural network for targeted sentiment analysis. *Knowl.-Based Syst.* **2020**, *188*, 104827. [CrossRef]
50. Haralabopoulos, G.; Anagnostopoulos, I.; McAuley, D. Ensemble deep learning for multilabel binary classification of user-generated content. *Algorithms* **2020**, *13*, 83. [CrossRef]
51. Liu, M.; Chen, L.; Du, X.; Jin, L.; Shang, M. Activated gradients for deep neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**.[CrossRef]
52. Irie, K.; Schlag, I.; Csordás, R.; Schmidhuber, J. Going Beyond Linear Transformers with Recurrent Fast Weight Programmers. *arXiv* **2021**, arXiv:2106.06295.
53. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
54. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
55. Hua, Y. Understanding BERT performance in propaganda analysis. *arXiv* **2019**, arXiv:1911.04525.
56. Topal, M.O.; Bas, A.; van Heerden, I. Exploring transformers in natural language generation: Gpt, bert, and xlnet. *arXiv* **2021**, arXiv:2102.08036.
57. Golestani, M.; Razavi, S.Z.; Borhanifard, Z.; Tahmasebian, F.; Faili, H. Using BERT Encoding and Sentence-Level Language Model for Sentence Ordering. In Proceedings of the International Conference on Text, Speech, and Dialogue, Olomouc, Czech Republic, 6–9 September 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 318–330.
58. Gu, X.; Liu, L.; Yu, H.; Li, J.; Chen, C.; Han, J. On the transformer growth for progressive bert training. *arXiv* **2020**, arXiv:2010.12562.

59. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv* **2019**, arXiv:1910.01108.
60. Alzahrani, E.; Jololian, L. How Different Text-preprocessing Techniques Using The BERT Model Affect The Gender Profiling of Authors. *arXiv* **2021**, arXiv:2109.13890.
61. González-Carvajal, S.; Garrido-Merchán, E.C. Comparing BERT against traditional machine learning text classification. *arXiv* **2020**, arXiv:2005.13012.
62. 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. *J. Med. Internet Res.* **2021**, *23*, e25431. [CrossRef]
63. Twitter Rules and Policies. Available online: https://help.twitter.com/en/rules-and-policies#twitter-rules (accessed on 1 November 2021).
64. List of Terrorist Incidents in 2021. Available online: https://en.wikipedia.org/wiki/List_of_terrorist_incidents_in_2021 (accessed on 1 November 2021).