2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE) | 978-1-6654-9260-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/IITCEE57236.2023.10091025

Predicting Abnormal User Behaviour Patterns in Social Media Platforms based on Process Mining

Sharanya G
Department of Computer Science and Engineering
SRM Institute of Science and Technology
Chennai, India
saranyag@srmist.edu.in

Deepanjali Chandrasekaran
Department of Computer Science and Engineering
SRM Institute of Science and Technology
Chennai, India
deepuchandru2001@gmail.com

Muli Dipikha Sre
Department of Computer Science and Engineering
SRM Institute of Science and Technology
Chennai, India
dipikhasremuli@gmail.com

Mithileysh Sathiyanarayanan
Research & Innovation
MIT Square, London
City, University of London
London, UK
mithileysh@mitsquare.com

***Abstract*— Cyberbullying has been one of the adverse repercussions of social media these days. The intensity of cyberbullying is risen considerably as a due to the increased use of image sharing and textual comments. Cyberbullying can be defined as transmitting, publishing, or circulating unbearable, harmful, false, or cruel content about some other individual. To keep the site safe and secure, automated processes for detecting certain instances are now vital. Process mining seems to be a combination of methodologies that integrate data science with management system to aid in the evaluation of organizational functions using log information. Whenever images and language that appear to be benign are combined, they might send bullying texts. As a result, different methods for analysing text and photos may fail to detect all instances of cyberbullying. In this study, we attempted to detect various examples of cyberbullying by combining textual data. The proposed system discovers hidden connections between individuals and members of a group who have similar behaviours. We achieve this through a novel strategy that incorporates techniques including data mining, analyzation of business procedures, and much more. Our study summarizes complete methodological process involved in the proposed system from the computer-generated data source in which the data pertaining user behavior, actions, and processes in an OS, software, website, or other sources to provide the visual representation of the abnormalities. Moreover, phase of identifying user behavioral patterns is the one which deserves attention.**

Keywords—Cyber Bulling, Cybercrime, social media, Process Mining, Data Mining.

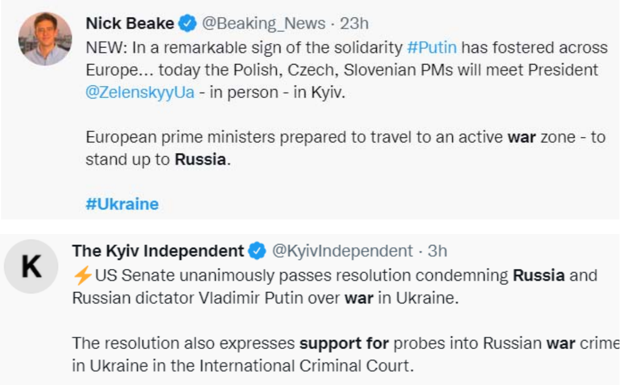
I. Introduction

Criminals and terrorists are not unique to the internet world, however the present popularity of social media has intensified the involvement inside the social web.[1] Spreading confidential and sensitive information over the internet, prompting insensitive comments, is an example of cyberbullying. This can result in complications for the individual regarding of revealing personal or confidential information on them. Many types of harassment are restricted or prohibited. Developmental disorders, increasing tension and worry, unhappiness, react back violently, and less self-confidence all are adverse impacts of cyberbullying. Regardless, the harassment is subsided, harassment might have long-term emotional consequences.

The Natural Language Toolkit (NLTK) is a Python high - level language used in quantitative language processing to interface on textual data (NLP). It includes text processing programmes for lemmatization, parser, categorisation, stemming, tagging, and semantics reasoning. Furthermore it comes with a recipe book and even a guide which outlines the ethics behind such core language processing techniques with which NLTK provides, along with visual examples and different sampling data sets. TensorFlow is a free, open-source machine learning tool. It's a set of tools for symbolic mathematics one which integrates dataflow with differentiable computing can handle wide range of tasks relevant to deep neural network training and testing phases. This enables programmers could design ML models via the wide range of tools, modules, and open-source resources. Criminal activists often use social networking sites to incite hatred against nations in order to achieve a certain goal. Russia-Ukraine war (2022) is also an example to create activists for and against the country. As the name implies, cyberbullying is use of online to harass individuals who are known or unknown towards the offender.

Cyberbullying had already resulted in various severe consequences for people affected, spanning from outbursts of rage to suicidal ideation. It's vital to note that content on social media would be both textual and visual. Text classification has maintained a popular area of research during the past decade. This is widely used for a variety of things, including sentiment analysis, brand analysis, collaborative filtering, and news article categorization into games, politics, and finance. Overall textual content of social media is indeed the focus of this study. The process of identifying dangerous content inside a textual information is just a purpose of classifying, and sentiment analysis or text categorization are most popular techniques [12]. The main aim of this proposed system is to perform real time cyberbullying analysis on the tweets that are extracted from the twitter and provide time-based analytics to the user.

978-1-6654-9260-7/23/$31.00 ©2023 IEEE



Nick Beake @Beaking_News · 23h

NEW: In a remarkable sign of the solidarity #Putin has fostered across Europe... today the Polish, Czech, Slovenian PMs will meet President @ZelenskyyUa - in person - in Kyiv.

European prime ministers prepared to travel to an active **war** zone - to stand up to **Russia**.

#Ukraine

The Kyiv Independent @KyivIndependent · 3h

⚡US Senate unanimously passes resolution condemning **Russia** and Russian dictator Vladimir Putin over **war** in Ukraine.

The resolution also expresses **support for** probes into Russian **war** crime in Ukraine in the International Criminal Court.

Fig. 1. Social Media contents to support or against the Russia-Ukraine War

II. RELATED WORKS

The link for prediction model combines metadata with a real-time crime database to better understand real-world circumstances and enhance link prediction quality. According to a survey of relevant work, the majority of the systems are built using traditional ML techniques like support vector machine (SVM) without using metadata integration. Improved resiliency and durability, as well as increasing the problem's complexities [2].

The goal of a WordNet-based criminal information mining system is to find and retrieve forensic investigation necessary data from massive suspicious recorded conversations. The structural characteristics an accused's communication log to determine a group of clusters and the themes which each cluster discussed. The method's efficacy and adaptability are further demonstrated by the findings. Thus, it reduces computational dimensionality, unreliability, as well as inefficiency [3].

Within area of computer vision, automatic human activity recognition (HAR) via computing equipment is a difficult problem to solve. With the development of smart gadgets such as cellphones, sensor systems like the accelerometer and gyroscope could utilized in order to monitor the daily physical motions with ease. For these applications, state-of-the-art deep neural network models such as Convolutional Neural Network (CNN) are sufficient. Traditional approaches need a significant amount of time to deploy and could be done in live time [4].

For two criminal behaviors that entail medium to extensive use of mobile platforms, bullying and poor drug dealing, the mixed criminal profile and suspect pattern recognition technique has indeed been developed. The best technique is chosen, as well as the situations are re-run on such a real dataset for more validation and evaluation. Improvements in operational excellence and
remedies have been found to be ineffectual [5].

A combinatorial event descriptive model was given based on source paper evaluations and determination of the aspects of criminal occurrences with the corresponding constituents. The schema allows for the methodical combination of numerous elements or criminal features. To achieve superior performance, performed best in a variety of situations, environments, and challenges [6].

Consumer drones are appeared recently as a concern in a variety of real-world situations due to their difficulty in detecting and tracking them, and their ease of use for illegal behavior such as smuggling illegal substances, surveillance operations, system breaching and so on. Innovations already in use are either ineffective at detecting objects size is around10 cm or are prohibitively costly and difficult to install. Increase initial and maintenance expenses by dramatically increasing correlation intensity using finer and much more concise data [7].

Codetection is indeed a scam detector that simultaneously employs a graph-based similarity measures or a feature matrix. It proposes a revolutionary way for revealing features of economic transactions such as fraudulent tendencies and suspicious assets. The suggested technique (Codetection) is effective in identifying malicious activities and also questionable features, accordance with test findings of simulated as well as the real time datasets. Built-in error handling as well as a high level of installation and maintenance complexity [8].

The addition of time-frequency characteristics streamlines the process of selecting attributes and increases data science model's quality. For such rest time within ML model development for fraudulent financial identification, time-frequency characteristics such as standard deviation, mean and variance were utilized. In order to meet modern network business expectations, a well-balanced tradeoff between numerous characteristics must be achieved [9].

Crime Analyzer, which worked closely with subject specialists and put their findings into a visualization system, developed a visual information interpretation platform to aid in the investigation of crimes in localities. Descriptive and analytical analysed, in addition to instance examples based on available data and input from subject matter experts, were used to validate the system. Enhance time productivity by minimizing computational resources time, as well as the mass and complexities of its deployment [10].

A common storage engine for big data platforms is required to identify the data remnants of investigative significance via Sync, or personal cloud storage provider. The results of the study add to a deeper knowledge of cloud-enabled big data storage forensic, potentially saving time and money being involved in real life inquiries employing sync on any cloud systems. Conventional methods have a significant workload. Thus, this method is simple, quick, and less difficult but its uncontrolled, opportunist mechanism [11].

III. PROPOSED SYSTEM

The proposed system is an innovative automated process that predicts cyberbullying based on people's social media reviews and tweets. The below seem to be the features of suggested cyberbullying system:

1. Automatically categorise and classify cyberbullying tweets;

2. Offer insights on cyberbullying incidents and online trolls;

The proposed system analyses and categorises cyberbullying, conducts feature-based categorization, handles ambiguity, and summarises viewpoints. For estimating sentiment, an accurate technique is applied, which aids in the improvement of marketing efforts. There are few drawbacks of the existing system which includes difficulties in obtaining better performance, it cannot be accurately applied in the real time sources of data. It also requires high complexity of installing and maintaining as well as it is bit time-consuming.

The major aim of proposed solution is just to analyse tweets retrieved from Twitter in live time detecting cyberbullying and deliver time-based metrics to the user. As shown in the fig 2 the Sentiment analysis has been performed on the following levels:

1. Document Level

2. Sentence level

3. Entity /Aspect Level

Events could be generated in the developed framework by considering each table containing a datetime column belongs to an action, as well as the value throughout this field relate to the entity's occurrences. The proposed technique is based on unique networking site model, and with a focus on customer behaviour. The change will have an impact on a description of the user-to-user connection. The basic approach to social network analysis is based on the notion that what a social networking site is a bunch of individuals (or groups of individuals) who socially interact. Communications, individual knowledge of one another, companionship, organization, as well as other aspects of social connection are all examples of social media interactions. To explain the composition of items, meta modelling might be utilised for cyberbully concept. Furthermore, the processes focused on social media seem to be more dynamic, and that there are no rigorous or organised procedures to control individuals' behaviour; instead, declarative constraints or boundaries between activity are being used.

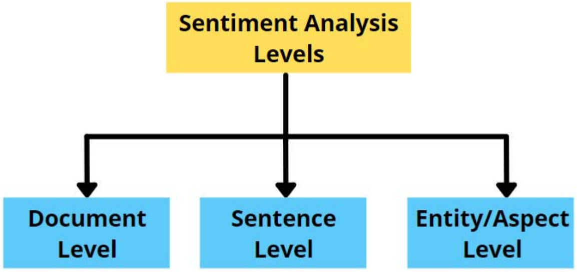


Fig. 2. Sentiment Analysis Level

IV. METHODOLOGY

A. Loading the data

This dataset comprises of information relating to the rapid recognition of cyber-bullying from multiple sources. The information was gathered from a variety of social channels, namely Kaggle, Twitter, Wikipedia Discussion pages, and YouTube. The dataset contains 3 attributes (Serial Number, Tweets, Text Label) and in total it has1064 records. Text mining is the practice of extracting meaningful information from textual data using natural language processing (NLP) tools and research techniques. Fig 3 shows the step-by-step process of text mining. Twitter data is a vast source of knowledge that may be utilized to acquire data on every topic imaginable. This information can be utilized for a variety of purposes, including identifying patterns relating to a given phrase, assessing brand awareness, and soliciting input on new goods / services. Twitter is an information treasure trove. Unlike the other social media platforms, virtually every individual's tweet was entirely public and retrievable. It is a significant benefit if you really need to collect a large dataset for analysis.

B. Data Preprocessing

Preprocessing raw data is indeed an essential part of creating a Schizophrenic Discourse model. Pre - processing stage includes a number of key procedures, including data cleansing, transformation of data, and extraction of features. Discourse in Schizophrenia Data cleaning and transformations are ways for removing outliers and standardizing information so it can be utilized to develop models more effectively.

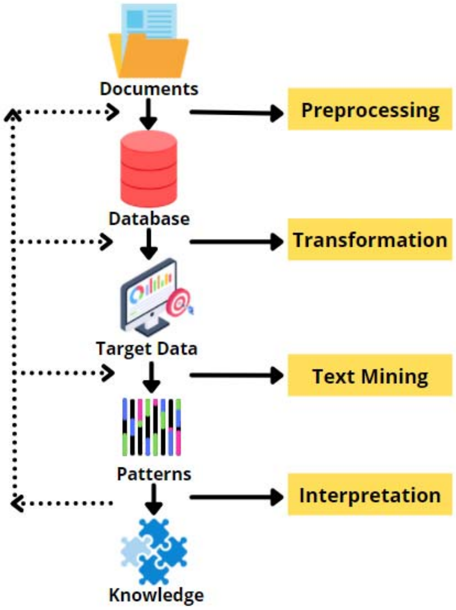


Fig. 3. Text Mining Process

Fig 4 shows the complete process of the proposed study. Hundreds of elements (descriptive terms) could be present in a Schizophrenic Discourse set of data; nevertheless, most of these categories will include redundancy. To reduce the model's complexity, those factors that provide distinctive and relevant data should be used. Attributes which do not add to a Schizophrenic Discourse framework can be eliminated using data mining methods.

As we retrieve and categorize features, we go through a variety of processing stages. The library is used to perform named entity (NE) recognition, coreference resolution, and dependency parsing.

The majority of discourse data is

- Noisy: Containing errors or anomalies

- Inconsistent: Having differences in code or identities
- Tasks in data preprocessing.
- Data Cleaning: Replace in incomplete data, clean inaccurate data, detect or eliminate outliers, and resolve conflicts
- Incomplete: Missing attribute values, missing certain key aspects, or simply providing publicly available data.
- Using numerous databases, data cubes, or files for data integration.
- Data transformations include data normalization and aggregation.
- Data Compression: Decreasing the amount of data while maintaining same or equivalent analytical outcomes.

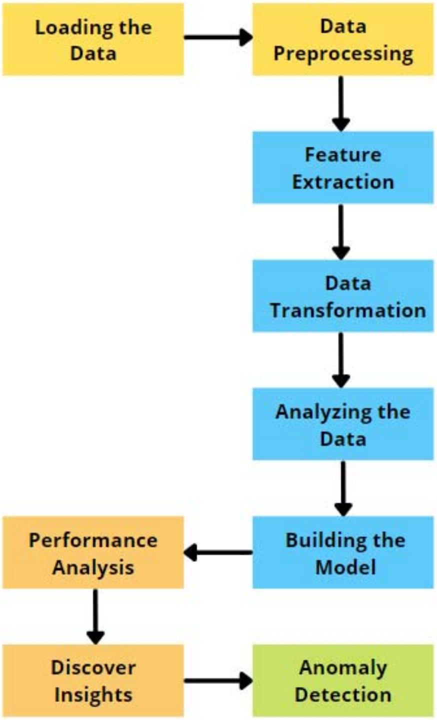


Fig. 4. Flowchart of proposed research method

- Data Discretization: Replacement of numeric qualities by nominal ones is the process of data reduction.

Many words appear often in the Schizophrenic Discourse dataset, despite the fact that they are largely meaningless since they were being used to combine phrases of terms. This is generally believed that stop words have no impact on the framework or contents of literary works. Depends on the the frequency with which it occurs, the textual existence information poses an impediment to evaluating all details of the information. Stop words like or, were, that, and some are frequently used. They're useless for classification tasks. As a result, these has to be eliminated. However, compiling a collection of stop words is hard and contradictory across different text documents. This approach will help our Schizophrenic Discourse method perform better by reducing textual information. Every textual data in the Schizophrenic Discourse series contains those words, which aren't required for text mining techniques.

C. *Building the Model*

A recurrent neural network is a form of network, which tries to represent timing or sequencing related behavior like languages, market values, or power demand. Is therefore done by fed directly the outcome of such a neural network layer during time t to same networking layer's source at time t + 1.

A recurrent neural network which makes use of cell blocks rather than the neural network layers in the earlier days. The input gate, forget gate, and output gate are three different components found in it. Fig 5 shows the process of textual data analyzation using ML and deep learning methods (NLTK and TensorFlow). With the help of various classifiers like Naïve Bayes, Decision Tree, Maxent and Support Vector Machine the textual data was analyzed. The storage enables a model to learn long term relationships among a series, letting that to forecast next word or phrase, an emotion categorization, or next temperature monitoring while considering the actual context. A network is designed to simulate how humans interpret sequences rather than responding to individual phrases, we examine the full sentence in developing an action.

N-grams seem to be continuous set of words, signs, or tokens. It could be described as the adjacent sequence of words in such text in technical language. While dealing with textual information in NLP (Natural Language Processing) procedures, they come into consideration.

V. RESULT AND DISCUSSION

Text mining process was involved to analyze the textual data from the dataset which was collected from twitter. Based on the analysis, the fig 6 shows that 624 non-bullying and 440 bullying text was identified from the data. It dramatically improves time productivity by minimizing calculation and communication times.

Table 1 shows the evaluation metrics (Precision, Recall and F-Measure) for all the classifiers in involved in the process of analyzation. A higher precision method computes quite effective findings over insignificant alternatives, while a high recall method delivers a majority of relevant data.

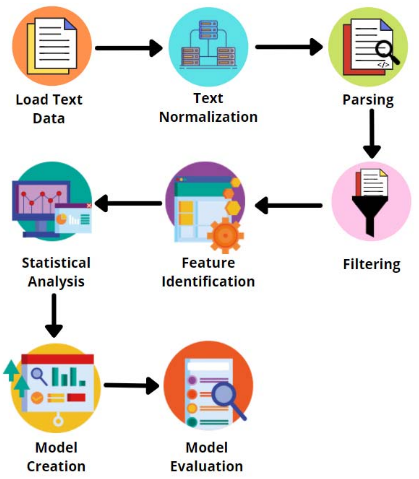


Fig. 5. Visual representation of proposed method

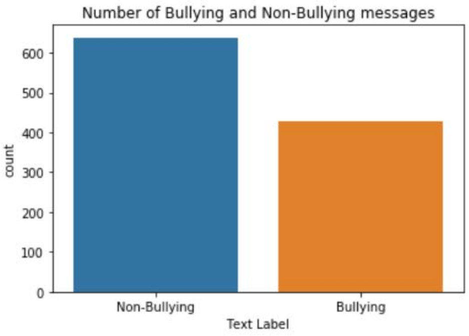
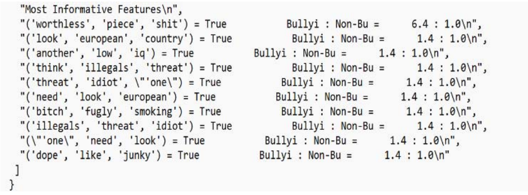


Fig. 6. Number of bullying and non-bullying texts in given dataset

TABLE I. EVALUATION METRIC VALUES FOR VARIOUS ALGORITHMS

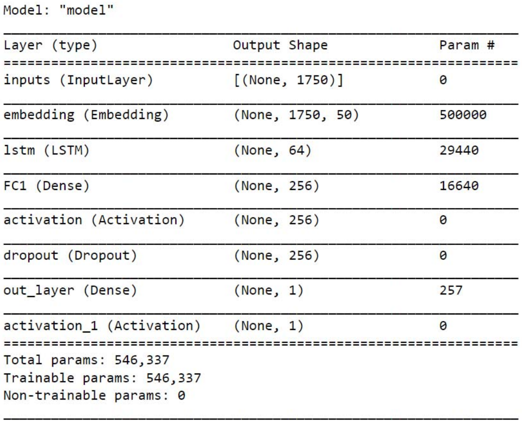
Evaluation metrics	Naïve Bayes Classifier	Decision Tree Classifier	Maxent Classifier	Support Vector Classifier
Bullying Precision	0.58	0.63	0.63	0.60
Bullying Recall	0.67	0.31	0.57	0.61
Bullying F-measure	0.62	0.41	0.6	0.61
Non - Bullying Precision	0.74	0.64	0.63	0.63
Non - Bullying Recall	0.66	0.87	0.93	0.93
Non - Bullying F-measure	0.70	0.74	0.75	0.75



```
  "Most Informative Features\n",
  "('worthless', 'piece', 'shit') = True           Bullyi : Non-Bu =      6.4 : 1.0\n",
  "('look', 'european', 'country') = True           Bullyi : Non-Bu =      1.4 : 1.0\n",
  "('another', 'low', 'iq') = True           Bullyi : Non-Bu =      1.4 : 1.0\n",
  "('think', 'illegals', 'threat') = True           Bullyi : Non-Bu =      1.4 : 1.0\n",
  "('threat', 'idiot', \"'one\") = True           Bullyi : Non-Bu =      1.4 : 1.0\n",
  "('need', 'look', 'european') = True           Bullyi : Non-Bu =      1.4 : 1.0\n",
  "('bitch', 'fugly', 'smoking') = True           Bullyi : Non-Bu =      1.4 : 1.0\n",
  "('illegals', 'threat', 'idiot') = True           Bullyi : Non-Bu =      1.4 : 1.0\n",
  "(\"'one\", 'need', 'look') = True           Bullyi : Non-Bu =      1.4 : 1.0\n",
  "('dope', 'like', 'junky') = True           Bullyi : Non-Bu =      1.4 : 1.0\n"
 ]
}
```

Fig. 7. Ratio between bullying and nonbullying for trigrams by using nltk

The dataset goes under the tokenization process and then words get jumbled randomly, from which the ratio of bullying and non-bullying for unigram, bigrams, trigrams and ngrams can be examined. The fig 7 shows the ratio of bullying and non-bullying for trigrams.



```
Model: "model"
_________________________________________________________________
Layer (type)                Output Shape              Param #
=================================================================
inputs (InputLayer)         [(None, 1750)]            0
_________________________________________________________________
embedding (Embedding)       (None, 1750, 50)          500000
_________________________________________________________________
lstm (LSTM)                 (None, 64)                29440
_________________________________________________________________
FC1 (Dense)                 (None, 256)               16640
_________________________________________________________________
activation (Activation)     (None, 256)               0
_________________________________________________________________
dropout (Dropout)           (None, 256)               0
_________________________________________________________________
out_layer (Dense)           (None, 1)                 257
_________________________________________________________________
activation_1 (Activation)   (None, 1)                 0
=================================================================
Total params: 546,337
Trainable params: 546,337
Non-trainable params: 0
_________________________________________________________________
```

Fig. 8. RNN Value for the Final Dataset

By supplying the identical weights and biases to all of the levels, RNN converts independent signals into dependent signals, minimizing the complication of rising factors and memorising every prior result by feeding every result into next hidden layer. The dataset was splitted up into different layers likes long short term memory (LSTM), dense, dropout, activation and so on. Fig 8 shows the various layers with output shape and the parameters (Trainable and Non-trainable parameters).

VI. CONCLUSION

The proposed theory of cyberbullying Process Mining in this paper and the designed framework about the Content Derivative algorithm overcomes the drawbacks of the existing system. The purpose of process mining is to extract the implementation records of activities in order to locate, analyze, and enhance them. System which enables individuals to be more flexible in their behavior (systems without even a structured manner or unlimited rules on user behavior) provide log data where a huge number of events, frequently quite identical, can be extracted. In certain platforms, the introduced mechanism enables for the unambiguous depiction of underlying relationships between individuals (or groups of users) having similar actions. This was one of the elements that led to the discovery of cyberbullying user's behavioral patterns. The second cause would have been to identify relevant forms of user behavior, or behavior patterns, to properly characterize the systems from such broader perspective.

As a result, the proposed method can be used in a variety of situations in real-world logs as well as more complicated processes. According to the precision value, Naïve Bayes algorithm is healthier and performs better than the other algorithms. The complete analyzation of the test set showed the accuracy of 0.725 with a loss 0.544. In future, we hope to expand the same cyberbullying approach with content derivative algorithm towards many other social media platforms for reducing threats and will improve the research work to involve combination of visual data and textual data to have more accurate evidence for cyberbullying crime activities [13-20].

REFERENCES

[1] I. Shafi, S. Din, Z. Hussain, I. Ashraf and G. S. Choi, "Adaptable Reduced-Complexity Approach Based on State Vector Machine for Identification of Criminal Activists on Social Media," in *IEEE Access*, vol. 9, pp. 95456-95468, 2021.

[2] M. Lim, A. Abdullah, N. Jhanjhi and M. Khurram Khan, "Situation-Aware Deep Reinforcement Learning Link Prediction Model for Evolving Criminal Networks," in *IEEE Access*, vol. 8, pp. 16550-16559, 2020.

[3] F. Iqbal, B. C. M. Fung, M. Debbabi, R. Batool and A. Marrington, "Wordnet-Based Criminal Networks Mining for Cybercrime Investigation," in IEEE Access, vol. 7, pp. 22740-22755, 2019.

[4] R. Mondal, D. Mukherjee, P. K. Singh, V. Bhateja and R. Sarkar, "A New Framework for Smartphone Sensor-Based Human Activity Recognition Using Graph Neural Network," in *IEEE Sensors Journal*, vol. 21, no. 10, pp. 11461-11468, 15 May15, 2021.

[5] K. Barmpatsalou, T. Cruz, E. Monteiro and P. Simoes, "Mobile Forensic Data Analysis: Suspicious Pattern Detection in Mobile Evidence," in *IEEE Access*, vol. 6, pp. 59705-59727, 2018

[6] G. Tsakalidis and K. Vergidis, "A Systematic Approach Toward Description and Classification of Cybercrime Incidents," in *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 4, pp. 710-729, April 2019.

[7] G. Casasanta *et al*., "Consumer Drones Targeting by Sodar (Acoustic Radar)," in *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 11, pp. 1692-1694, Nov. 2018.

[8] D. Huang, D. Mu, L. Yang and X. Cai, "CoDetect: Financial Fraud Detection With Anomaly Feature Detection," in *IEEE Access*, vol. 6, pp. 19161-19174, 2018

[9] U. G. Ketenci, T. Kurt, S. Önal, C. Erbil, S. Aktürkoğlu and H. Ş. İlhan, "A Time-Frequency Based Suspicious Activity Detection for Anti-Money Laundering," in *IEEE Access*, vol. 9, pp. 59957-59967, 2021.

[10] G. Garcıa *et al*., "CrimAnalyzer: Understanding Crime Patterns in São Paulo," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 4, pp. 2313-2328, 1 April 2021.

[11] Y. Teing, A. Dehghantanha, K. R. Choo, Z. Muda and M. T. Abdullah, "Greening Cloud-Enabled Big Data Storage Forensics: Syncany as a Case Study," in *IEEE Transactions on Sustainable Computing*, vol. 4, no. 2, pp. 204-216, 1 April-June 2019.

[12] M. Asif, A. Ishtiaq, H. Ahmad, H. Aljuaid, and J. Shah, ‘‘Sentiment analysis of extremism in social media from textual information,’’ Telematics Informat., vol. 48, May 2020, Art. no. 101345.

[13] M. Sathiyanarayanan, C. Turkay, and O. Fadahunsi. "Design of Small Multiples Matrix-based Visualisation to Understand E-mail Socio-organisational Relationships." In 2018 10th International Conference on Communication Systems & Networks (COMSNETS). 2017.

[14] M. Sathiyanarayanan,, and D. Pirozzi. "Spherule diagrams with graph for social network visualization." In 2016 8th International Conference on Communication Systems and Networks (COMSNETS), pp. 1-6. IEEE, 2016.

[15] M. Sathiyanarayanan,, and D. Pirozzi. "Social network visualization: Does partial edges affect user comprehension?." In 2017 9th international conference on communication systems and networks (COMSNETS), pp. 570-575. IEEE, 2017.

[16] M. Sathiyanarayanan, AK. Junejo, and O. Fadahunsi. "Visual Analysis of Predictive Policing to Improve Crime Investigation." In 2019 International Conference on contemporary Computing and Informatics (IC3I), pp. 197-203. IEEE, 2019.

[17] M. Sathiyanarayanan, and C. Turkay. "Is multi-perspective visualisation recommended for e-discovery email investigations?." (2016).

[18] M. Sathiyanarayanan, and C. Turkay. "Determining and Visualising E-mail Subsets to Support E-discovery." (2016).

[19] M. Sathiyanarayanan, and O. Fadahunsi. "Integrating Digital Forensics and Digital Discovery to Improve E-mail Communication Analysis in Organisations." In Smart Computing Paradigms: New Progresses and Challenges, pp. 187-193. Springer, Singapore, 2020.

[20] M. Sathiyanarayanan, C. Turkay, and O. Fadahunsi. "Design and implementation of small multiples matrix-based visualisation to monitor and compare email socio-organisational relationships." In 2018 10th International Conference on Communication Systems & Networks (COMSNETS), pp. 643-648. IEEE, 2018.