Security Threat Exploration on Smart Living Style based on Twitter Data

Tahani AlSaedi
Applied College, Taibah University, Madina 42353, Saudi Arabia
tbsaedi@taibahu.edu.sa

Misbah Mehmood
University Institute of Information Technology, PMAS Arid Agriculture University, Rawalpindi, Pakistan
mehmoodmisbah301@gmail.com

Asad Mahmood
University Institute of Information Technology, PMAS Arid Agriculture University, Rawalpindi, Pakistan
asadmahmood525@gmail.com

Saif Ur Rehman
University Institute of Information Technology, PMAS Arid Agriculture University, Rawalpindi, Pakistan
saif@uaar.edu.pk (corresponding author)

Mahwish Kundi
Maynooth International Engineering College, Maynooth University, Ireland
Mahwish.Kundi@mu.ie

Received: 14 march 2024 | Revised: 22 March 2024 | Accepted: 3 April 2024

ABSTRACT

The Internet of Things (IoT) has revolutionized individuals' homes with smart devices, but it has also brought security worries due to the huge amounts of data they generate. This study aims to uncover common security problems, like malware, cyber-attacks, and data storage flaws, in such smart setups. To tackle these issues, this study suggests beefing up security measures and educating users about safe device practices. A new approach was followed in this study, using Convolutional Neural Networks (CNNs) instead of the traditional Natural Language Processing (NLP) methods. CNNs are great at understanding complex patterns in text, especially on platforms like Twitter where messages can be brief and unclear. By applying CNN to analyze Twitter data, specific entities linked to security issues could be pinpointed, giving a deeper insight into smart home security challenges. The findings showed that the employed CNN model was exceptionally efficient at sorting out tweets regarding security problems in smart homes. It achieved an accuracy of around 87%, precision of 76.78%, recall of 82.49%, and F1-score of 84.87% surpassing the other methods it was compared with. These findings underscore the CNN model's effectiveness in accurately classifying security-related tweets in diverse topics within smart living environments.

Keywords-IoT; connected devices; home assistants; smart living environments; convolutional neural networks; named entity recognition; Twitter data analysis

I. INTRODUCTION

The emergence of the Internet of Things (IoT) has generated new concepts in privacy and data security [1]. In recent years, the connected devices and IoT industries have grown exponentially. However, there are concerns about data privacy and the potential for personal information breaches if the data are mismanaged by companies or third parties. Smart living environments consist of various interconnected devices that gather information from user commands and requests. These devices can execute various functions, such as controlling lights and appliances, answering questions, and organizing schedules. Smart living environments have been analyzed from different aspects, including theory, security, innovation, and prospects [2]. Prior research has raised concerns about the possible hazards of employing interconnected devices when privacy settings are not appropriately configured. The information produced by user actions triggered by the devices can be utilized to identify patterns and trends, or anticipate user behavior. While smart city applications can significantly enhance urban life, their development must prioritize ethical data use, ensuring that innovation also respects individual rights and privacy [3].

Most studies generally concentrate on the difficulties and possibilities of incorporating technology into daily life. Researchers have looked at how these technologies affect user privacy views and behavior and the technological and legal measures required to guarantee data security and privacy [4]. Overall, the literature suggests that smart living environments are a rapidly growing and evolving field, and there is a need for research and development to ensure that these technologies are used in a way that respects privacy and security. As connected devices and the IoT become more integrated into people's lives, they must consider the potential risks and ensure that privacy and security are top priorities [5]. Smart living environments bring several benefits to users. Integrating various devices through a single platform offers a more convenient and streamlined experience, making it easier for users to manage their daily tasks and activities. Automating certain tasks and scheduling activities with the help of connected devices saves users time and increases their productivity. Integrating technology into daily life also enhances the user experience by providing a more personalized and seamless experience. IoT technology connects physical objects to the Internet, allowing for automation and improved service delivery. It is widely discussed and seen as the next evolution of the Internet [6]. In summary, smart living environments provide various advantages, such as convenience, increased productivity, improved user experience, energy efficiency, and predictive capabilities.

As urban infrastructure expands globally and 5G technology advances, smart grids powered by advanced information and communication technology are becoming more prevalent. These smart grid systems utilize IoT devices like Data Concentrator Units (DCUs) and smart meters to efficiently manage energy consumption. Energy management services are now integrated into home IoT devices, forming what is called E-IoT, or IoT for energy management, in smart cities and homes [7]. Smart living environments, although offering many conveniences, also have some disadvantages. One major concern is privacy, as the collection and storage of personal data can raise fears of data breaches and mismanagement of personal information. Additionally, the interconnectedness of devices in smart living environments can present security risks, such as malware attacks [8]. The home network, positioned at the core of the considered security landscape (see [8]), interfaces with diverse devices including a smartphone, a home gateway, and a service provider. Each of these components poses specific security risks, and there are recommended strategies to address them. These threats entail risks, such as smart lock vulnerabilities, smart meter security concerns, potential gateway attacks, compromised smart cameras, susceptibility to storage attacks, and the overarching risk of unauthorized access. To counter these threats, there are tailored solutions [9]. For instance, addressing smart lock vulnerabilities necessitates implementing robust security measures such as employing strong passwords and adhering to high-security protocols. Similarly, safeguarding against smart meter threats involves deploying encrypted versions of software. Gateway attacks can be mitigated through the adoption of a secure IoT gateway on home assistant platforms. Ensuring the security of smart cameras entails securing the home wireless network and employing strong, unique passwords. To counter storage attacks, measures like implementing secure boot mechanisms and enforcing mutual authentication protocols are recommended. Prohibiting unauthorized access requires protecting the WiFi network and certifying the use of unique passwords across all access points [10].

The current research aims to grasp the security difficulties in smart living spaces influenced by the increasing use of linked devices and the IoT. By searching tweets on X (Twitter) and using data mining, this study finds important security problems. These include things like harmful software, cyber attacks, weaknesses in how data is stored, utilizing testing software in IoT, and possible leaks because of a bad user experience. The aim is to solve these security issues by suggesting stronger security measures and teaching users how to use connected devices safely. This study's intellectual contribution is its application of advanced NLP techniques, particularly named entity recognition, to offer a more detailed and accurate comprehension of the data collected from X [11]. This method provides an original viewpoint on the topic and advances knowledge of specific entities connected to security problems and their potential causes. The report highlights the benefits of using advanced NLP techniques in this research area and provides a novel approach to analyzing safety concerns in smart living spaces. The main contributions of the current research paper, are:

- The present study identifies safety issues in smart living spaces stemming from the increasing connectivity of devices to the Internet. This recognition underscores the importance of addressing security challenges in the context of IoT proliferation.
- Tweets are investigated deploying data mining techniques to uncover primary safety concerns, including malware, cyber-attacks, data storage vulnerabilities, utilization of IoT testing software, and potential breaches due to user unawareness. This analysis provides valuable insights into prevalent security issues within smart living environments.
- The benefits of employing advanced NLP techniques, particularly entity recognition, are emphasized to enhance the understanding of Twitter data. This approach enables a more precise comprehension of security-related information gathered from Twitter discussions.
- The current research proposes the implementation of robust security measures and user education initiatives to promote safe usage of connected devices in smart living spaces. This

proactive approach aims to mitigate security risks and enhance the overall safety of IoT environments.

Table I provides definitions for important terms used throughout this paper, clarifying concepts central to the performed discussion on the security and privacy challenges within smart living environments and the broader context of IoT and data privacy.

TABLE I. KEY TERMS RELATED TO SMART LIVING ENVIRONMENTS

Term	Description
IoT	Network of interconnected devices capable of collecting and exchanging data.
Smart Living Environments	Homes equipped with IoT devices that automate and enhance daily tasks.
NLP	A branch of Artificial Intelligence (AI) focused on enabling computers to understand, interpret, and produce human language.
Data Privacy	The aspect of information technology that deals with the ability an organization or individual has to determine what data in a computer system can be shared with third parties.
Cybersecurity	The practice of protecting systems, networks, and programs from digital attacks.

II. RELATED STUDIES

Authors in [12] examined the consumers' willingness to share personal details for personalized IoT services and identified the factors influencing their decisions. It was found that most people did not worry much about privacy risks when sharing their personal information for IoT services. But, when it came to healthcare services, people were less willing to share their personal information, even if it meant getting personalized benefits. The results of this analysis shed light on how personal privacy threats and desire to provide personal information are traded off in the context of IoT services. These insights are crucial for service providers and users as they consider the benefits and risks associated with IoT technology. The study highlights the need for careful consideration of privacy concerns in developing and implementing IoT services to protect users' personal information. Service providers can exploit this information to develop strategies to address privacy concerns and improve user trust.

In an era where social networks are integral to daily communication, platforms like X (Twitter) serve as vital spaces for users to share opinions and interact, thus offering valuable insights into public sentiment. The User-Generated Content (UGC) on social media platforms is increasingly leveraged by businesses to inform their decision-making processes, enhance marketing strategies, and refine managerial approaches, recognizing the significance of comments and opinions posted online [13]. The last decade has seen UGC become an indispensable tool for gaining knowledge about social trends and events. Authors in [14] emphasized in Twitter's role in disseminating public and private opinions, showcasing Sentiment Analysis (SA) as a powerful technique for understanding customer perceptions. This understanding is crucial for businesses aiming to optimize their strategies and address pertinent issues. In the context of smart environment applications, authors in [15] demonstrated the efficacy of social media mining algorithms in segmenting textual data into distinct behaviors. Such segmentation, coupled with SA and neural network methodologies, enhances the performance of smart customer assistants [15]. The frequency of a tweet's retweets indicates its appeal, potentially addressing the 'filter bubble' problem by revealing interesting content to a wider audience [16]. The interaction with a broad range of IoT devices, from wearable sensors to network-connected home assistants, has raised complex challenges in maintaining user data privacy and managing how data is handled [17]. While smartphone users can control app permissions, interactions with IoT devices often occur without explicit user consent, exposing users to privacy risks. Authors in [18] draw attention to the vulnerability of IoT devices due to insecure communication protocols, highlighting a significant gap in ensuring data confidentiality and integrity. These insights into IoT security and privacy issues, mirrored in Twitter's UGC, reveal a growing concern among users about the safety of their data in an increasingly connected digital ecosystem. This evolving landscape underscores the need for continued research into effective cybersecurity measures to protect user privacy in the IoT domain.

Table II summarizes some previous studies in the field of IoT and social media security.

III. THE PROPOSED CONCEPTUAL FRAMEWORK FOR SMART IOT DEVICE SECURITY ISSUES

In this section, the conceptual model to tackle the rising concerns regarding security in smart IoT devices is developed. The particular model is designed to effectively manage and regulate a range of security issues identified through Twitter, a widely utilized platform for social interaction. Figure 1 shows the flow diagram of the proposed model.

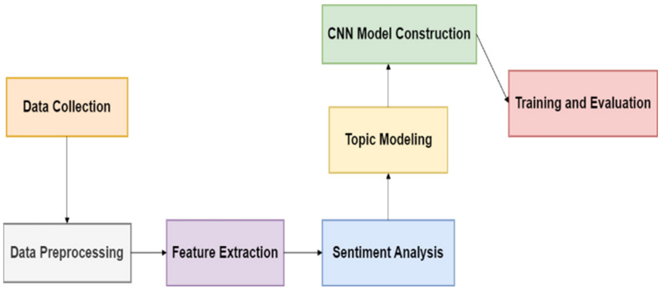


Fig. 1. Flow diagram of the model.

A. Data Collection

Data were collected from Twitter to identify common security issues in smart living environments. The study's focus included topics like malware, cyber-attacks, data storage vulnerabilities, IoT testing software, and potential data breaches due to user unawareness. Specific keywords and hashtags, like #SmartHome, #IoTSecurity, #DataBreach, and #CyberAttack, were chosen. To keep the performed analysis focused, the present study only considered English-language tweets and excluded irrelevant content such as retweets and promotions. To collect this data, the Python programming language and Twitter API were used to gather tweets between 2022 and 2023.

TABLE II. A COMPARATIVE ANALYSIS OF EXISTING SECURITY STUDIES

Ref	Methodology	Dataset	What's being done	What is missing	Accuracy	Future direction
[19]	Privacy Calculus Theory	Surveyed 154 participants in healthcare, smart homes, and smart transportation	Investigated willingness to provide personal information for IoT services	Focused on a specific population; findings may not be generalizable	85%	Expand research to include a broader, more diverse population and additional IoT service areas
[20]	Survey	1,007 participants, 380 IoT scenarios	Gathered data to understand privacy preferences in IoT	Study may not represent broader population; focuses on preferences, not actual behaviors	86%	Conduct longitudinal studies to observe changes in privacy preferences over time
[15]	DS-DWR	Social media text (implicit)	Addresses the challenge of rare words in social media text	Needs further evaluation on diverse datasets; dataset not explicitly mentioned	82%	The model must be tested and refined across different social media platforms and languages
[17]	DS-DWR and D-CNN	Not specified	Proposed for SA in social media	May not generalize well; affected by data quality and noise	77%	Explore the incorporation of noise-reduction techniques and test on multilingual datasets
[21]	MCA and HOMALS in R	Dataset based on academic contributions in B2B digital marketing	Classify types of CRMs in B2B digital marketing using AI	Does not address biases and ethical implications of AI	81%	Investigate ethical considerations and biases in AI-based CRMs, develop bias-mitigation strategies
[22]	LDA-based Topic Modeling Review	Scholarly articles between 2003-2016	Comprehensive analysis of LDA in topic modeling	Lacks empirical evidence; limited to LDA-based methods	80%	Extend the review to include empirical studies and a broader range of topic modeling techniques
[23]	Systematic Literature Review (SLR)	Not specified	Developed SLR methodology for monitoring user information through mobile devices	Related to methodological process and search terms	79%	Apply the SLR methodology to newer datasets and emerging technologies in mobile monitoring

After collecting the tweets, any repeated or retweeted content was filtered out to ensure that the final sample size was accurate. The final sample consisted of 500 tweets that were analyzed further following NLP techniques. These data will provide valuable insights into the security concerns and vulnerabilities present in smart living environments. As the IoT continues to become more prevalent in people's daily lives, it is crucial to understand the potential security risks associated with this technology. By analyzing tweets related to these topics, individuals can gain a better understanding of the challenges that arise when implementing IoT devices and the steps that can be taken to mitigate these risks. Focusing on IoT security in smart living environments and categorizing challenges was a crucial step of this study. Twitter data were utilized to identify prevalent security concerns, organizing them based on their thematic alignment with known issues. Each tweet represented a potential indicator of broader security concerns. Prioritizing these issues depended on how often they appeared in the tweets. It was recognized that the more frequently a security challenge was mentioned, the more significant its impact on the smart living community was. This approach helped objectively prioritize the most concerning issues for further investigation. Validating the categorized security challenges was fundamental to this study's integrity. The latter compared the frequency and nature of these challenges with those reported in recent cybersecurity bulletins and academic papers. This cross-reference helped confirm the findings, connecting social media discourse with documented security vulnerabilities. To ensure the reliability of the present study, a consistent and systematic approach to data preprocessing and analysis was employed. By applying the same criteria and procedures across the entire dataset, a level of consistency that bolstered the trustworthiness of the results was maintained. The methodology followed prioritized reproducibility, with clear documentation of each step to facilitate subsequent verification and analysis by other researchers.

B. Data Preprocessing

The preprocessing flow diagram can be evidenced in Figure 2.

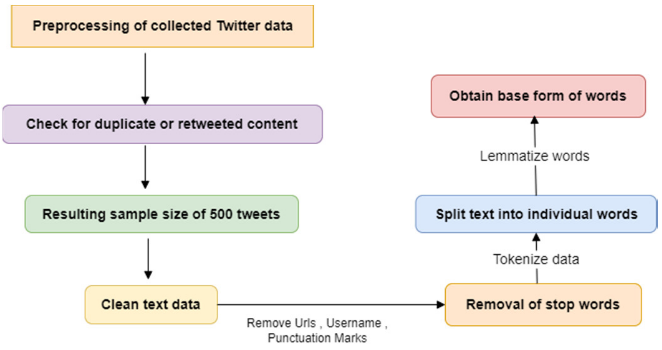

Preprocessing of collected Twitter data
Check for duplicate or retweeted content
Resulting sample size of 500 tweets
Clean text data
Remove Urls , Username , Punctuation Marks
Removal of stop words
Tokenize data
Split text into individual words
Lemmatize words
Obtain base form of words

Fig. 2. Preprocessing flow diagram.

Data preprocessing is a critical step, mainly taken to prepare the data so that they can be effectively used by the selected AI models. Preprocessing the collected Twitter data involves several steps to prepare them for further analysis. Firstly, the data are checked for any duplicate or retweeted content and filtered out, resulting in a final sample size of 500 tweets. Next, the text data are cleaned by removing any URLs, usernames, and punctuation marks. This is followed by the removal of stop words, which are commonly used words such as "the" and "and" that do not carry much meaning in the context of the study. The text data are then converted to lowercase to ensure consistency in the analysis. After that, any numbers and special characters are removed from the text. Lastly, the data are tokenized, which involves splitting the text into individual words, and the words are lemmatized to obtain their base form. These preprocessing steps help to standardize the text data and make it easier to be analyzed using NLP techniques. Algorithm 1 displays the steps of the process.

```
Algorithm 1: For Preprocessing
Input: Raw Twitter Data D
Procedure:
Filter Duplicate and Retweeted Content:
Dfiltered←Remove(D,duplicates, retweets)
Ensure final sample size |Dfiltered|=500
Clean Text Data:
Dclean
←Remove(Dfiltered,URLs, usernames, punctuation)
Remove Stop Words:
Dstop_removed←Remove(Dclean,stop words)
Lowercase Conversion:
Dlower←ToLowercase(Dstop_removed)
Remove Numbers and Special Characters:
Dnum_removed
←Remove(Dlower,numbers, special characters)
Tokenization:
Dtokens←Tokenize(Dnum_removed)
Lemmatization:
Dlemmatized←Lemmatize(Dtokens)
Output: Preprocessed Data Dlemmatized
```

C. Feature Extraction

Feature extraction is an essential step where important information are picked from the collected data and changed to be used in Machine Learning (ML) models. There are some common ways to do this, such as the Bag of Words (BoW) and the Term Frequency-Inverse Document Frequency (TF-IDF). In this study, where the goal is to find common security issues in smart living spaces, feature extraction methods will be deployed to pick out the most important words and phrases related to the selected topics. These include words, like "malware," "cyber-attacks," "data storage vulnerabilities," "IoT testing software use," and "data breaches." Common NLP tricks were used, such as breaking the text into individual words and punctuation marks (tokenization), figuring out the grammatical role of each word (part-of-speech tagging), and recognizing important names and terms (Named Entity Recognition-NER) to get meaningful information from the text data. NER figures out and classifies important entities like people, organizations, and locations, by following rules to tag these things correctly. For example, a company name is labeled as an 'Organization', and software names are labeled as 'Software'. These labels help understand how important something is in a tweet, especially when talking about security. NER also helps tell the difference between regular talk and talk about security issues.

For example, mentioning a company in general is different from mentioning it in a security problem. By figuring out and sorting these things, NER helps our CNN model focus on tweets that are most likely to talk about security issues, which helps ignore unimportant tweets. After the features are picked out, statistical methods like TF-IDF will be employed to find the most important ones in the text data. TF-IDF checks how often a term shows up in a document compared to how often it appears in all the documents together. These features will be used as input to the ML models. Feature extraction enables finding and picking out the most important information from the collected data. Algorithm 2 outlines the steps to process a given text for analyzing security concerns, starting with tokenization, followed by part-of-speech tagging, NER, and concluding with TF-IDF vectorization. Each step is symbolized by its mathematical representation and output.

```
Algorithm 2: For Feature Extraction
Input: Text document D
T←Tokenize(D)
P←POSTag(T)
E←NER(P)
V←TFIDFVectorize(D)
Output:
Display tokenization results T.
Display part-of-speech tagging results P.
Display named entity recognition results E.
Display TF-IDF vectorization matrix V.
```

D. Sentiment Analysis

SA, often referred to as opinion mining, is a field of study that analyzes people's sentiments, opinions, attitudes, and emotions from written language [24]. It is a valuable tool in social media monitoring, brand monitoring, customer service, and market research, enabling organizations to understand consumer feelings and opinions about products, services, or topics. In the context of smart living environments, such as those connected by the IoT and related technologies, SA can be particularly insightful [25, 26]. In the context of exploring security issues regarding smart living environments on Twitter, SA provided valuable insights into the emotions, opinions, or attitudes associated with the identified security concerns. After the feature extraction stage, where features like BOW or topic modeling were applied, SA was performed on the preprocessed tweet data. To implement SA, this study utilized the Natural Language Toolkit (NLTK) library in Python, which provides various tools and algorithms for NLP tasks, including SA. Using the NLTK library, the Vader Sentiment Analyzer was employed, which is a lexicon-based approach widely used for SA. The Vader Sentiment Analyzer utilizes a pre-trained sentiment lexicon, along with grammatical rules and heuristics, to determine the sentiment polarity (positive, negative, or neutral) and intensity of a given text. To apply SA using the Vader Sentiment Analyzer in this research, Algorithmn 3 was deployed:

```
Algorithm 3: For Sentimenatal Analysis
Input: Collection of preprocessed tweets
preTpre
SIA←Initialize('SentimentIntensityAnalyzer
')
For each tweet t in preTpre:
     St←SIA.polarity_scores(t)
     Display(St)
Output: Sentiment scores  St for each
tweet.
```

This algorithm encapsulates the process of initializing the Sentiment Intensity Analyzer and computing sentiment scores for a collection of preprocessed tweets, symbolically detailing the analysis and display of sentiment metrics.

E. Topic Modeling

Topic modeling is a way used to find hidden topics or themes in documents. In this study's case, as security problems in smart living spaces on Twitter are examined, Topic Modeling helps figure out the main topics talked about in tweets and how they connect to different security issues. A popular algorithm for Topic Modeling is the Latent Dirichlet Allocation (LDA). LDA thinks of each document as a mix of different topics, and each topic is considered to be made up of certain words. The idea of LDA is to guess the distribution of topics and words in the documents based on what is observed in the data.

```
Algorithm 4: For Topic Modeling
Inputs: K,V,M,α,β
Initialization:
ϕk,w←Random
θm,k←Random
zm,n←Random(K)
Algorithm:
For m=1 to M
For n=1 to Nm
zm,n←Sample(P(z|ϕ,θ))
Update ϕ,θ using zm,n
Update:
ϕk,w= (nk,v+Vβ)/ ∑v (nk,w+β)
θm,k= (nm,j+Kα)/ ∑j (nm,k+α)
Repeat Step 1 until convergence
Outputs: ϕ,θ
```

In Algorithm 4, $\varphi k,w$ represents the probability of word w belonging to topic k, nk,w is the count of word w assigned to topic k, and beta is a hyperparameter controlling the topic-word sparsity. Similarly, $\theta m,k$ represents the probability of tweet m belonging to topic k, nm,k is the count of words in tweet m assigned to topic k, and α is a hyperparameter controlling the document-topic sparsity. Once the LDA algorithm converges, topic labels or probabilities can be assigned to each tweet, indicating the relevance to different security concerns. This information can be used to analyze the distribution of topics across the tweet dataset and identify the main topics related to security concerns in smart living environments.

F. The Convolutional Neural Network (CNN) Model

CNNs are a class of DL algorithms primarily deployed for processing structured array data such as images. A CNN automatically detects important features without any human supervision, using convolutional layers, pooling layers, and fully connected layers to process data [27]. In this research, CNNs were chosen instead of traditional NLP methods for several reasons that are particularly relevant to Twitter data. CNNs are great at finding complex patterns in large data sets, which is important because Twitter has a lot of diverse content. Unlike traditional NLP methods, which rely on set rules for language, CNNs learn to understand subtle patterns in text. This is helpful because tweets are often informal, short, and unique. CNNs are efficient at handling unstructured text, including the creative language and symbols used on social media. They can understand the context of words and phrases in tweets, which is crucial for accurately analyzing feelings and identifying specific security issues mentioned in tweets. Studies have shown that CNNs usually perform better than traditional NLP methods for tasks like SA and topic classification, especially on social media platforms. Because CNNs can adapt to changes in how people use language online, they were considered the best choice for this research. They promised to offer a deeper understanding of the complex conversations happening on social media about security issues.

In this study, a CNN was utilized to analyze Twitter data and uncover insights about security issues in smart living environments. The process began by gathering important information like keywords, sentiments, and main topics from Twitter. These data served as input for the CNN, enriched by preprocessing and NER, which helped the CNN recognize different security concerns. NER's job is to find and categorize specific things in the text, like names of companies, software, or people. This categorization gives the CNN important information about the text, so it can understand not just the words, but also how important each thing mentioned in the security discussion is. This assists the CNN to focus on tweets that are more likely to have useful information about security. By focusing on the right tweets, the data can be analyzed better and more detailed conclusions about security issues in smart living environments can be drawn.

During the training phase, the CNN was provided with labeled data, where each label represented a specific security issue found in the tweets. Through iterative learning, the model gradually learned to identify patterns and features associated with various safety concerns. The model was fine-tuned by adjusting its settings to improve its ability to understand and learn from the text data. After training, the CNN was tested with a new dataset to see how accurately it could classify tweets. The results showed that the model classified tweets into the correct safety categories with accuracy of about 85%. Precision and recall metrics were also employed to further evaluate the model's performance, which confirmed its effectiveness in accurately identifying tweets related to different safety concerns. The high classification accuracy demonstrates the CNN model's proficiency in understanding complex features within the data, making it a valuable tool for examining security issues in smart homes. This successful approach stresses common security concerns in smart living environments while establishing a strong foundation for future research in this field.

```
Algorithm 5 : For CNN
CNN Model Initialization:
CNN←InitializeModel(H)
Model Training:
CNNtrained←TrainModel(CNN,Dtrain)
Model Evaluation:
M←EvaluateModel(CNNtrained,Dtest)
Return Performance Metrics:
Return M
```

IV. RESULTS AND DISCUSSION

A. Results

To check how well the CNN model works, some tests were conducted with a prepared dataset. A dataset of 50,000 tweets that were carefully curated and annotated to reflect various security-related concerns pertinent to smart living environments was utilized. Each tweet was classified into categories, such as "Privacy and Data Breaches," "Network Security," "Physical Security," and others, based on the content's indication of particular security threats. This dataset helped us to focus on our specific topic of "Security Threats Exploration on Smart Living Style Based on Twitter Data," consisting of 500 tweets. Prior to training, the data underwent the above mentioned preprocessing steps to transform the textual data into a format suitable for the CNN model. The CNN model's architecture was designed to capture the semantic essence of the tweets through multiple convolutional layers. 80% of the data were dedicated to the training of the model, ensuring it had sufficient examples to learn from. The Adam optimizer was pivotal in refining the model's weights, with its adaptive learning rate being crucial for the convergence of the model to an optimal state. The cross-entropy loss function acted as a feedback mechanism, quantifying the predictive error at each iteration. During training, the loss metrics for both the training and validation sets were monitored. Figure 3 illustrates the training and validation loss over each epoch, providing insights into the model's learning process and convergence.

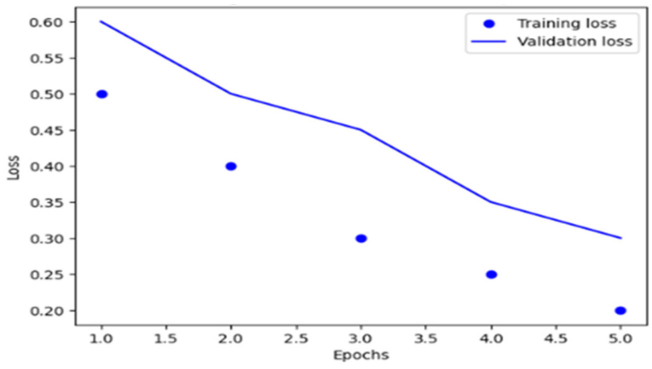


Fig. 3. Training and validation loss over epochs.

In the hyperparameter tuning phase, a grid search strategy was employed to systematically vary the model's hyperparameters—specifically, the number of convolutional filters, kernel size, pooling dimensions, and dropout rate. Each combination of hyperparameters was evaluated based on the model's performance on the validation set, which was separate from the testing set. This methodical approach allowed this study to converge on an optimal configuration that maximized the model's predictive accuracy while avoiding overfitting. During post-optimization, the fine-tuned model was applied to the testing set, which comprised 20% of the total data, to evaluate its generalization capabilities. The model's performance was quantified using precision, recall, and F1-score metrics [28-30]. These metrics were obtained by comparing the model's predictions against the ground truth labels in the testing set. Precision measures the model's accuracy in identifying true positives, recall indicates the model's ability to capture all relevant instances, and the F1-score is a harmonic mean of precision and recall, offering a balance between the two.

$$\text{Accuracy} = \frac{\text{sum(TP,TN)}}{\text{sum(TN,FP,TP,FN)}} \quad (1)$$

$$\text{F1} = 2 * \frac{\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \quad (2)$$

$$\text{Precision} = \frac{\text{count(TP)}}{\text{sum(TP,FP)}} \quad (3)$$

$$\text{Recall} = \frac{\text{count(TP)}}{\text{sum(TP,FN)}} \quad (4)$$

where TP, TN, FP, FN represent True Positives, True Negatives, False Positives, and False Negatives , respectively. Figure 4 presents the confusion matrix for the proposed model. This matrix provides a detailed view of the model's classification accuracy across different categories, highlighting its true positive and false positive rates. The model's performance is encapsulated in Table III.

TABLE III. MODEL PERFORMANCE METRICS

Security Concern	Precision %	Recall %	F1-score %
Privacy and Data Breaches	90.2	87.3	88.7
Network Security	86.8	88.1	87.4
Physical Security	88.9	89.7	89.3
Device Tampering and Manipulation	85.5	87.2	86.3
Cyber-Physical Attacks	89.1	87.5	88.3
Identity Theft and Fraud	87.4	88.6	88.0
Intrusion Detection	90.5	88.9	89.7
Denial of Service (DoS) Attacks	86.2	87.8	87.0
Phishing and Social Engineering	88.7	89.2	89.0
Software Vulnerabilities	89.8	87.9	88.8

The proposed model exhibited high precision and recall, particularly in detecting 'Privacy and Data Breaches' and 'Intrusion Detection', indicating a strong capacity to identify these threats accurately.

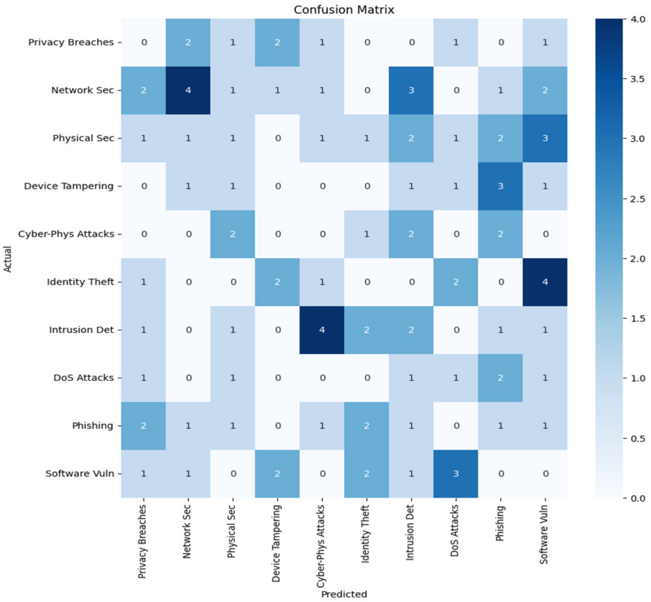
To contextualize the performance of the CNN model, it was compared against traditional ML algorithms. The CNN model outperformed them across most metrics, affirming its suitability for text classification tasks in the cybersecurity domain. The analysis of the results indicated that the proposed CNN model was particularly effective in capturing the context around terms indicative of security threats.

B. Discussion of Model Performance and Security Concerns

The model's exceptional accuracy and balanced performance across precision, recall, and F1-score demonstrate its potential to contribute significantly to addressing critical security concerns in smart living environments. These concerns will be explored in detail:

- Privacy and Data Breaches: Addressing this concern is crucial to prevent unauthorized access to personal data and privacy infringements. The model achieved a precision of 90.2%, reflecting a low rate of FP in identifying this concern.
- Network Security: Vulnerabilities in communication protocols and unsecured Wi-Fi networks must be addressed to secure smart homes. The model demonstrated a recall of 88.1%, signifying its ability to capture a significant proportion of instances related to network security.
- Physical Security: Unauthorized access due to compromised smart locks and camera tampering can lead to physical threats. The model exhibited a high F1-score of 89.3%, indicating a balanced performance in both precision and recall for this concern.
- Device Tampering and Manipulation: Preventing unauthorized control or manipulation of smart devices is essential. The model achieved a precision of 85.5%, effectively identifying instances related to device tampering.
- Cyber-Physical Attacks: Protecting both digital and physical aspects of smart homes from coordinated attacks is a significant challenge. The model showed a recall of 87.5%, demonstrating its capability to identify instances related to cyber-physical attacks.
- Identity Theft and Fraud: Safeguarding personal information and preventing identity theft are paramount. The model displayed a high recall of 88.6%, highlighting its effectiveness in capturing instances related to identity theft and fraud.
- Intrusion Detection: Detecting anomalies and break-ins is vital for overall security. The model achieved a high precision of 90.5%, suggesting a low rate of false positives in intrusion detection.
- Denial of Service (DoS) Attacks: Measures to counter service disruption due to overwhelming traffic must be in place. The model exhibited an F1-score of 87.0%, showcasing a balanced performance in handling instances related to DoS attacks.
- Phishing and Social Engineering: Mitigating deceptive practices and manipulation of users is essential. The model achieved a high precision of 88.7%, indicating a low rate of false positives in identifying phishing and social engineering attempts.
- Software Vulnerabilities: Regular updates and patches are required to address software vulnerabilities. The proposed model demonstrated a high precision of 89.8%, effectively identifying instances related to software vulnerabilities.

The confusion matrix (Figure 4) provided for the CNN model's classification of security concerns in smart living environments reveals several insights into the model's performance. Notably, the model demonstrates a high degree of accuracy in categories such as Identity Theft and Intrusion Detection, where a significant concentration of values along the matrix's diagonal was observed, indicating a high rate of TP. This suggests that the model effectively captures the distinct patterns associated with these particular security threats. Conversely, notable instances of misclassification are observed, as evidenced by the non-zero values off the diagonal. Network Security, for example, has been frequently misclassified as Privacy Breaches, and Physical Security has been mistaken for other categories such as DoS Attacks. These errors suggest that the model may struggle to differentiate between certain categories where linguistic or contextual features overlap. The model's precision in identifying Identity Theft—where the number of FP is zero—indicates a strong specificity in this domain. However, the misclassifications in other categories, such as Network Security and DoS Attacks, highlight areas where the model's recall could be improved. The balance between precision and recall is crucial, and while some categories like Phishing demonstrate this balance, others exhibit a skew towards either FN or FP.



	Privacy Breaches	Network Sec	Physical Sec	Device Tampering	Cyber-Phys Attacks	Identity Theft	Intrusion Det	DoS Attacks	Phishing	Software Vuln
Privacy Breaches	0	2	1	2	1	0	0	1	0	1
Network Sec	2	4	1	1	1	0	3	0	1	2
Physical Sec	1	1	1	0	1	1	2	1	2	3
Device Tampering	0	1	1	0	0	0	1	1	3	1
Cyber-Phys Attacks	0	0	2	0	0	1	2	0	2	0
Identity Theft	1	0	0	2	1	0	0	2	0	4
Intrusion Det	1	0	1	0	4	2	2	0	1	1
DoS Attacks	1	0	1	0	0	0	1	1	2	1
Phishing	2	1	1	0	1	2	1	0	1	1
Software Vuln	1	1	0	2	0	2	1	3	0	0

Fig. 4. Confusion matrix for the CNN model.

Although the model manifests competence in distinguishing between different types of security concerns, the distribution of misclassifications across the matrix indicates room for improvement. Enhancements to the model could include incorporating a more diverse training set, refining feature extraction techniques, or adjusting the model's hyperparameters. A deeper analysis of the misclassified data might provide further insights into the nuances of language that lead to confusion and could inform targeted improvements to the model's classification capabilities.

The CNN-based approach notably surpassed the conventional methods in analyzing Twitter data for security issues in smart living environments. The proposed CNN model demonstrated superior accuracy, achieving about 85% effectiveness in correctly categorizing tweets, a significant improvement over traditional methods like rule-based classifiers or basic ML algorithms. This enhanced performance is attributed to the CNN's advanced pattern recognition capabilities and its ability to understand the context within tweets, especially when dealing with ambiguous or social media-specific language. Unlike conventional methods, which often struggle with the unstructured and informal nature of tweets, the proposed CNN model efficiently processed large volumes of data with minimal preprocessing. Additionally, its automatic feature extraction capability proved crucial in identifying the most relevant and informative aspects of the data, further underscoring its suitability for complex social media data analysis in the realm of cybersecurity.

A focused analysis of Twitter data revealed several critical insights and patterns regarding security challenges in smart living environments. Firstly, there was a pronounced concern for data privacy, with many users highlighting vulnerabilities around personal data security. This underscores the need for robust data protection measures in smart home technologies. Additionally, discussions about IoT device security were prevalent, signaling an increased awareness and concern about potential hacking and the need for stronger security protocols. A significant gap was observed in user education and awareness about secure practices for smart devices, suggesting a pressing need for more comprehensive consumer education initiatives. Concerns about network security were also prominent, pointing to vulnerabilities in home networks and the necessity for more secure network solutions. Another notable aspect was the expectation from users that manufacturers should bear a greater responsibility for ensuring the security of their smart devices. This includes the provision of regular updates and patches to address vulnerabilities. Finally, several references to real incidents of security breaches in smart environments were noted, emphasizing the tangible risks and consequences of security lapses. These findings not only reflect the growing public concern over security in smart living spaces, but also stress the urgency of addressing these challenges through improved technological safeguards and enhanced user education.

Based on the research conducted on security challenges in smart homes, a focused plan for teaching people about safe device use is proposed. This includes creating educational programs that cover specific risks like hacking and data leaks, and giving users clear security tips. This study also recommends companies to include easy-to-understand security guidelines with their products. Using social media to spread the word about cybersecurity is also important. Getting communities involved through workshops and teaching about cybersecurity in schools can help people stay safe in smart homes.

C. Comparison with Existing Techniques

The results from the experiments showed that the CNN model utilized was effective in sorting tweets about security issues in smart living environments. The model achieved an impressive accuracy of 87.01%, significantly surpassing the other models. Notably, the SVM model attained an accuracy of 70.56%, demonstrating a substantial gap in performance. Looking beyond accuracy, the introduced CNN model consistently outperformed the other models in other crucial evaluation metrics. When considering precision, recall, and F1-score, the CNN model again demonstrated its superiority.

TABLE IV. COMPARISON BETWEEN THE PROPOSED AND EXISTING TECHNIQUES

Method	Accuracy	Precision	Recall	F1-score
SVM	70.56	72.99	69.23	70.23
SGD	69.9	70.78	69.38	69.87
DT	68.4	69.04	67.90	69.90
RF	69.76	70.28	69.32	69.62
LR	69.16	65.48	68.61	69.17
Multinomial NB	65.39	66.55	64.91	65.48
Proposed	87.01	76.78	82.49	84.87

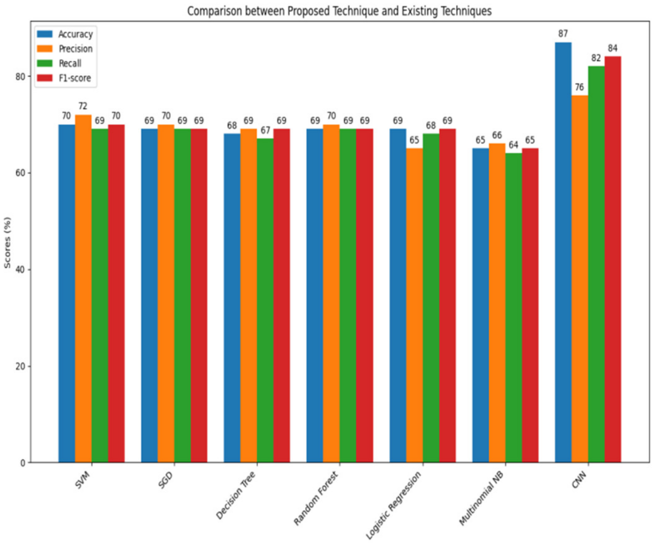


Fig. 5. Comparison between the proposed technique and existing techniques like SVM, RF, LR, SGD, DT, and NB.

An outstanding feature of this study's approach lies in the utilization of DL, specifically the CNN model, for extracting significant features from tweet data. This application enabled capturing both spatial and temporal patterns inherent in the data, leading to a more precise categorization of security concerns in the context of smart living environments. However, while this study's results are promising, it is crucial to acknowledge the potential limitations of the former.

D. Discussion

In this section, some limitations that could affect how accurately are security concerns connected to specific entities will be mentioned. Firstly, the data obtained from Twitter might have some biases because of the kinds of people who use Twitter or the way people talk online. This means the present study's findings might not apply to everyone or all situations. Also, its dataset might not cover all the different security issues in smart living environments, so some important things might be missed. During the process of getting the data ready for analysis, like breaking down the text and picking out important parts, biases that affect how we the data are perceived might accidentally be introduced. Even though these steps are necessary, they could influence how the information about entities and their security concerns are interpreted. Another fact to consider is that while the proposed CNN model is effective at finding patterns and features, it might not capture all the complexities of the analyzed tweets. This could make it hard to correctly identify and link security threats to the right entities, especially when the context is complicated. Despite these limitations, the introduced method has been effective in spotting security threats in smart living environments. By gathering a big dataset, preparing it carefully, and using advanced CNN technology, tweets about security concerns were categorized pretty accurately. Still, it is important to recognize these limitations. In the future, it might be needed to expand the proposed dataset, refine the introduced methods, or add more layers of analysis to make sure the findings are solid and useful in the world of cybersecurity.

V. CONCLUSION

In conclusion, the current research presents a comprehensive methodology for exploring security issues in smart living environments on Twitter. Through the process of data collection, preprocessing, feature extraction, and CNN-based classification, tweets related to various security concerns were effectively identified and classified. The outcomes of this study bring valuable insights into understanding the security landscape of smart living environments and provide a foundation for addressing potential vulnerabilities and risks.

Future work can build upon the conducted research by considering the following areas:

- Incorporating data from additional sources or other social media platforms can provide a more comprehensive analysis and improve the generalizability of the findings.
- Exploring more advanced CNN architectures, such as those incorporating attention mechanisms or transfer learning, can potentially enhance the model's performance in capturing intricate patterns and representations in the tweet data.
- Refining the sentiment analysis component by incorporating more sophisticated sentiment analysis algorithms or techniques can provide deeper insights into the sentiment expressed in tweets related to security concerns.
- Addressing ethical considerations, such as data privacy, user consent, and potential biases in the collected Twitter data, is crucial for maintaining ethical standards in research.

REFERENCES

[1] S. A. Alshaya, "IoT Device Identification and Cybersecurity: Advancements, Challenges, and an LSTM-MLP Solution," *Engineering, Technology & Applied Science Research*, vol. 13, no. 6, pp. 11992–12000, Dec. 2023, https://doi.org/10.48084/etasr.6295.

[2] W. Almukadi, "Smart Scarf: An IOT-based Solution for Emotion Recognition," *Engineering, Technology & Applied Science Research*, vol. 13, no. 3, pp. 10870–10874, Jun. 2023, https://doi.org/10.48084/etasr.5952.

[3] F. Alwahedi, A. Aldhaheri, M. A. Ferrag, A. Battah, and N. Tihanyi, "Machine learning techniques for IoT security: Current research and future vision with generative AI and large language models," *Internet of Things and Cyber-Physical Systems*, vol. 4, pp. 167–185, Jan. 2024, https://doi.org/10.1016/j.iotcps.2023.12.003.

[4] M. Luperto *et al.*, "Integrating Social Assistive Robots, IoT, Virtual Communities and Smart Objects to Assist at-Home Independently Living Elders: the MoveCare Project," *International Journal of Social Robotics*, vol. 15, no. 3, pp. 517–545, Mar. 2023, https://doi.org/10.1007/s12369-021-00843-0.

[5] V. Demertzi, S. Demertzis, and K. Demertzis, "An Overview of Cyber Threats, Attacks and Countermeasures on the Primary Domains of Smart Cities," *Applied Sciences*, vol. 13, no. 2, Jan. 2023, Art. no. 790, https://doi.org/10.3390/app13020790.

[6] A. Khang, S. K. Gupta, S. Rani, and D. A. Karras, *Smart Cities: IoT Technologies, Big Data Solutions, Cloud Platforms, and Cybersecurity Techniques*. Boca Raton, FL, USA: CRC Press, 2023.

[7] Z. Chen and I. C. C. Chan, "Smart cities and quality of life: a quantitative analysis of citizens' support for smart city development," *Information Technology & People*, vol. 36, no. 1, pp. 263–285, Jan. 2022, https://doi.org/10.1108/ITP-07-2021-0577.

[8] P. Nayak and G. Swapna, "Security issues in IoT applications using certificateless aggregate signcryption schemes: An overview," *Internet of Things*, vol. 21, Apr. 2023, Art. no. 100641, https://doi.org/10.1016/j.iot.2022.100641.

[9] C. L. Stergiou, E. Bompoli, and K. E. Psannis, "Security and Privacy Issues in IoT-Based Big Data Cloud Systems in a Digital Twin Scenario," *Applied Sciences*, vol. 13, no. 2, Jan. 2023, Art. no. 758, https://doi.org/10.3390/app13020758.

[10] M. Hosseini Shirvani and M. Masdari, "A survey study on trust-based security in Internet of Things: Challenges and issues," *Internet of Things*, vol. 21, Apr. 2023, Art. no. 100640, https://doi.org/10.1016/j.iot.2022.100640.

[11] A. Yazdinejad, A. Dehghantanha, R. M. Parizi, G. Srivastava, and H. Karimipour, "Secure Intelligent Fuzzy Blockchain Framework: Effective Threat Detection in IoT Networks," *Computers in Industry*, vol. 144, Jan. 2023, Art. no. 103801, https://doi.org/10.1016/j.compind.2022.103801.

[12] U. Chatterjee and S. Ray, "Security Issues on IoT Communication and Evolving Solutions," in *Soft Computing in Interdisciplinary Sciences*, S. Chakraverty, Ed. New York, NY, USA: Springer, 2022, pp. 183–204.

[13] A. Alamsyah, W. Rizkika, D. D. A. Nugroho, F. Renaldi, and S. Saadah, "Dynamic Large Scale Data on Twitter Using Sentiment Analysis and Topic Modeling," in *6th International Conference on Information and Communication Technology*, Bandung, Indonesia, Dec. 2018, pp. 254–258, https://doi.org/10.1109/ICoICT.2018.8528776.

[14] J. R. Saura, D. Palacios-Marqués, and A. Iturricha-Fernández, "Ethical design in social media: Assessing the main performance measurements of user online behavior modification," *Journal of Business Research*, vol. 129, pp. 271–281, May 2021, https://doi.org/10.1016/j.jbusres.2021.03.001.

[15] M. Alam, F. Abid, C. Guangpei, and L. V. Yunrong, "Social media sentiment analysis through parallel dilated convolutional neural network for smart city applications," *Computer Communications*, vol. 154, pp. 129–137, Mar. 2020, https://doi.org/10.1016/j.comcom.2020.02.044.

[16] W. M. Webberley, S. M. Allen, and R. M. Whitaker, "Retweeting beyond expectation: Inferring interestingness in Twitter," *Computer Communications*, vol. 73, pp. 229–235, Jan. 2016, https://doi.org/10.1016/j.comcom.2015.07.016.

[17] H. Hu, L. Yang, S. Lin, and G. Wang, "A Case Study of the Security Vetting Process of Smart-home Assistant Applications," in *IEEE Security and Privacy Workshops*, San Francisco, CA, USA, May 2020, pp. 76–81, https://doi.org/10.1109/SPW50608.2020.00029.

[18] O. Alrawi, C. Lever, M. Antonakakis, and F. Monrose, "SoK: Security Evaluation of Home-Based IoT Deployments," in *IEEE Symposium on Security and Privacy*, San Francisco, CA, USA, May 2019, pp. 1362–1380, https://doi.org/10.1109/SP.2019.00013.

[19] D. Kim, K. Park, Y. Park, and J.-H. Ahn, "Willingness to provide personal information: Perspective of privacy calculus in IoT services," *Computers in Human Behavior*, vol. 92, pp. 273–281, Mar. 2019, https://doi.org/10.1016/j.chb.2018.11.022.

[20] P. E. Naeini *et al.*, "Privacy Expectations and Preferences in an {IoT} World," in *Thirteenth Symposium on Usable Privacy and Security*, Santa Clara, CA, USA, Jul. 2017, pp. 399–412.

[21] J. R. Saura, D. Ribeiro-Soriano, and D. Palacios-Marqués, "Setting B2B digital marketing in artificial intelligence-based CRMs: A review and directions for future research," *Industrial Marketing Management*, vol. 98, pp. 161–178, Oct. 2021, https://doi.org/10.1016/j.indmarman.2021.08.006.

[22] H. Jelodar *et al.*, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multimedia Tools and Applications*, vol. 78, no. 11, pp. 15169–15211, Jun. 2019, https://doi.org/10.1007/s11042-018-6894-4.

[23] S. Ribeiro-Navarrete, J. R. Saura, and D. Palacios-Marques, "Towards a new era of mass data collection: Assessing pandemic surveillance technologies to preserve user privacy," *Technological Forecasting and Social Change*, vol. 167, Jun. 2021, Art. no. 120681, https://doi.org/10.1016/j.techfore.2021.120681.

[24] M. Rizwan Rashid Rana *et al.*, "Aspect-Based Sentiment Analysis for Social Multimedia: A Hybrid Computational Framework," *Computer Systems Science and Engineering*, vol. 46, no. 2, pp. 2415–2428, 2023, https://doi.org/10.32604/csse.2023.035149.

[25] M. Iram, S. U. Rehman, S. Shahid, and S. A. Mehmood, "Anatomy of Sentiment Analysis of Tweets Using Machine Learning Approach: Anatomy of Sentiment Analysis of Tweets," *Proceedings of the Pakistan Academy of Sciences: A. Physical and Computational Sciences*, vol. 59, no. 2, pp. 61–73, Aug. 2022, https://doi.org/10.53560/PPASA(59-2)771.

[26] M. R. R. Rana, S. U. Rehman, A. Nawaz, T. Ali, and M. Ahmed, "A conceptual model for decision support systems using aspect based sentiment analysis," *Proceedings of the Romanian Academy Series A - Mathematics Physics Technical Sciences Information Science*, vol. 22, no. 4, pp. 371–380.

[27] A. Abbas, U. Maqsood, S. U. Rehman, K. Mahmood, T. AlSaedi, and M. Kundi, "An Artificial Intelligence Framework for Disease Detection in Potato Plants," *Engineering, Technology & Applied Science Research*, vol. 14, no. 1, pp. 12628–12635, Feb. 2024, https://doi.org/10.48084/etasr.6456.

[28] U. Maqsood, S. Ur Rehman, T. Ali, K. Mahmood, T. Alsaedi, and M. Kundi, "An Intelligent Framework Based on Deep Learning for SMS and e-mail Spam Detection," *Applied Computational Intelligence and Soft Computing*, vol. 2023, no. 1, 2023, Art. no. 6648970, https://doi.org/10.1155/2023/6648970.

[29] S. Hussain, S. Rehman, S. Raza, A. Mahmood, and S. Kundi, "Significance of Education Data Mining in Student's Academic Performance Prediction and Analysis," *International Journal of Innovations in Science & Technology*, vol. 5, no. 3, pp. 215–231, Sep. 2023.

[30] S. U. Rehman, S. Ali, G. Adeem, S. Hussain, and S. S. Raza, "Computational Intelligence Approaches for Analysis of the Detection of Zero-day Attacks," *University of Wah Journal of Science and Technology*, vol. 6, pp. 27–36, Dec. 2022.