Date available online: 13-02-2025
Vol. 2, Issue 2 (June - December 2024)
This is an open-access article.

Auto-Classification of FIA-Cybercrime Wing Complaints Using Bidirectional Encoder Representations from Transformers Model

Hafiz Ammar Mazhar[1*], Umair Altaf[2], Syed Muhammad Anwar[1], Muhammad Shoaib Bhutta[4]**

[1]Department of Software Engineering, University of Engineering and Technology UET Taxila,Taxila, Pakistan
[2]University of Central Punjab, Lahore, Pakistan
[3]School of Automobile Engineering, Guilin University of Aerospace Technology, Guilin 541004, China

[*]Corresponding author: Hafiz Ammar Mazhar (e-mail: ammarmb@hotmail.com),
[**]Corresponding author: Umair Altaf (e-mail: umairaltaf8317@gmail.com).

Abstract— The Cybercrime Wing (CCW) of Federal Investigation Agency, which was formerly known as the National Response Center for Cybercrime (NR3C), is governed by rules that were created in 2016 as part of the Prevention of Electronic Crimes Act (PECA) to combat cybercrimes. Criminal activities executedusing computers and the internet are referred to as cybercrimes. In order to carry out illegal activities, cyber-criminals make use of any information system as their primary means of communication with the devices that belong to their victims. This research mainly focused on the cybercrime complaints with an automated classification system. To achieve the automatic modelling for classification of different types of cybercrimes, this study used the Bidirectional Encoder Representations from Transformers (BERT). Its obstacles mainly include the possibility of human errors as it manually classifies cybercrime complaints, also that there might be delays during handling in comparison with an automated system. The dataset includes complaints submitted in English during a two years window, and it was encoded, tokenized and cleaned thoroughly. The purpose was to simplify the training process for the model. The study used a lightlyfine-tuned, pre-trained (BERT)-base-uncased model. The findings confirm that the model can be used for classifying complaints and exhibits an excellent classification accuracy, precision and F1-scores between different cybercrime offenses indicating its supremacy among advanced Natural language processing (NLP)techniques to strengthen cybersecurity measures.

Index Terms—FIA, BERT, Cybercrime, Classification, NLP

I. INTRODUCTION

The cyber offenses are rapidly increasing, so the Cybercrime Wing (CCW) of Federal Investigation Agency is at Pakistan's forefront for its protection. Imposed under the Prevention of Electronic Crimes Act (PECA) 2016 [1], CCW includes a broad remit to investigate and prosecute cyber crimes. Equipped with the required instruments to counteract digital threats, ranging from financial fraud and online harassment, while ensuring the protection of digital rights, including privacy within the public at large. The formal establishment of CCW signifies a monumental step in Pakistan's drive against cybercrime, presenting firm determination of the state towards curbing this contemporary threat. However, the sheer quantity of complaints and complexity of cybercrimes necessitates creative options to boost the effectiveness & efficiency in CCW operations. Manual classification of cybercrime complaints has inherent problems, such as processing lag and an extended likelihood of human error. Collectively, these issues make CCW ill-equipped to respond quickly and efficiently to cyber threats, necessitating a transition towards automation rather than historic technology implementations. The release of an implementation for automated classification using the BERT model is a major step towards tackling the many problems in complaint classification [2].

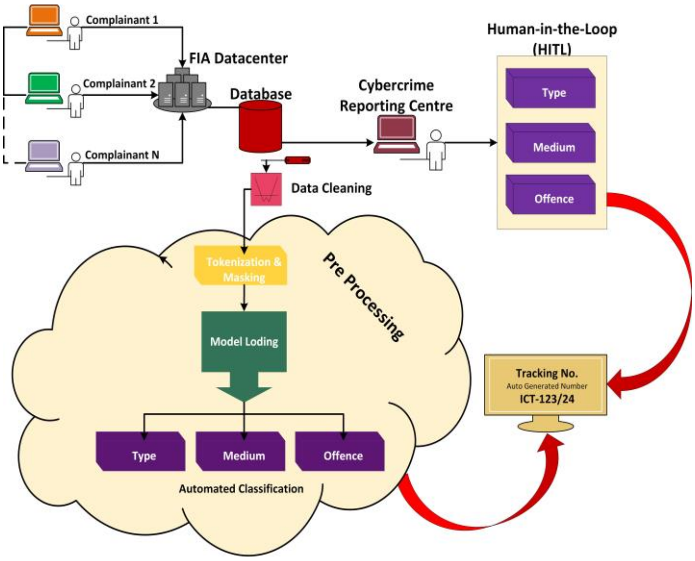

FIGURE 1. Automated Complaint Classification System Overview

In 2019, Kim [3] initially suggested a CNN-based text categorization model, which by utilizing Word2vec transforms a word into fixed-length vector and then employs multisite convolution to verify word-vector convolution. Pooling and categorization are the final steps. Convolutional neural network's primary benefit is its ability to extract local featured from text efficiently and fast training-speed. The pooling layer, however, will lose a substantial quantity of vital information and ignore the correlation between the whole and local. Wallace and Zhang [4] also suggested a CNN-based text classification approach and conducted multiple comparative experiments with varying hyperparameter settings. In addition, they provided guidance on parameter-tuning and had some experiments with hyperparameter configurations.

As BERT models have pushed the boundaries of text understanding by enhancing context-awareness through multilayer attention mechanisms, the proposed work by Faheem and Al-Khasawneh [5] pushes the boundaries of cyberattack detection in IoBC systems. Their use of deep learning to analyze complex data flows addresses a similar challenge of extracting meaningful patterns from vast and dynamic datasets.

The Deep Pyramid Convolutional Neural Network (DPCNN), proposed by Johnson and Tong [6], enhances text categorization by increasing the network depth to capture long-distance relationships in text. However, their computational demands limit their real-world applications. Similarly, the Graph-Based Convolutional Neural Network (GCN) proposed by Yao et al. [7] excels in classifying small datasets but faces challenges in terms of scalability and adaptability.

Aligns with the challenges of layered architectures discussed in deep learning models like CNNs, RNNs, and BERT for classification tasks. It focuses on how fog computing enhances IoT applications by bringing computational resources closer to the network edge, improving latency and real-time processing, much like CNNs efficiently extract local features in text classification tasks proposed by Burhan et al. (2023) [8]. The study also highlights security concerns in fog-IoT environments, which mirrors challenges in maintaining data integrity and context in NLP models, such as pooling in CNNs that can lose key information.

Mikolov et al. [9] utilized Recurrent Neural Networks (RNNs) for text classification, which are capable of processing inputs of varying lengths but are prone to gradient issues affecting learning efficiency. Models like Recurrent Neural Networks (RNNs) can face "gradient issues". In a vanishing gradient, the adjustments become too small to make progress, leading to slow or stalled learning. In an exploding gradient, they grow too large, causing unstable and erratic training. Both issues make it hard for models to learn effectively, especially in complex tasks like language processing. To overcome RNN's limitations of RNNs with long sequences, Schmidhuber and Hochreiter developed Long Short-Term Memory networks (LSTMs) [10], which, despite their improved performance, require extensive computation owing to their complex structures and numerous parameters.

Chung et al. [11] introduced the Gated Recurrent Unit (GRU) model, which streamlines the LSTM architecture for better training efficiency, though it still struggles with parallel computation and gradient issues. Graves and Schmidhuber [12] advanced LSTM to a bidirectional form (BiLSTM), improving classification, but at the cost of increased complexity. Cao et al. [13] combined BiGRU with contextual understanding to effectively categorize Chinese text, offering simplicity and faster convergence. Li and Dong [14] integrated Convolutional Neural Networks (CNNs) with BiLSTM to enhance text feature extraction for classification.

Faheem and company introduced a blockchain-based framework designed to enhance the security and resilience of distributed renewable energy systems. [15] This framework employs blockchain technology to ensure transparency and immutability, thereby securing the data related to energy events and mitigating unauthorized access through smart contracts and cryptographic algorithms. Although primarily focused on energy management, the principles of data integrity and decentralized control are highly relevant to the field of cybercrime, particularly in the context of automatically classifying cybercrime complaints using BERT. By leveraging the strengths of blockchain in securing complaint data, it is possible to enhance the reliability and transparency of automated classification systems, ensuring that the integrity of the complaint logs is maintained.

The bidirectional encoder representations from transformers (BERT) model [16] was further classified using its two way reading capacity to better understand the text context. Researchers [17] fine-tuned BERT and compared it with other models, such as KNN and SVM, using various sequences, batch sizes, and hyperparameter tuning to achieve sentiment classification. Hyperparameter-tuning is the process of finding the best combination of these settings to improve the model's performance. For example, in this study, different learning rates and batch sizes were tested to see which produced the best results. Fine-tuning these parameters helps optimize the model's ability to learn from data without overfitting or underfitting. Finally, [18] discussed utilizing BERT embedding with a deep neural network for classifying cognitive domains, emphasizing precision, recall, and F-measure for a balanced evaluation against class imbalance.

The original structure of the BERT model utilized only the last layer for classification, neglecting the semantic insights from the lower layers. To remedy this, the BERT-MLF model was created by integrating BERT's full 12-layer architecturevia a CNN, but it still lacks dynamic weight assignment to semantic information across layers. The BERT-MLDFA model [19] addresses this by dynamically incorporating parameters from all BERT layers using a multilevel attention mechanism, optimizing the classification for similar content categories, and enhancing the discrimination of key semantic information. Further developments in sentiment analysis forWeibo text involve enriching word vectors with an external sentiment dictionary and combining BERT, BiLSTM, and attention mechanisms with a CNN for feature extraction [20].This enhances the classification accuracy.

For hate speech detection [21], researchers have optimized BERT training with fine-tuning and logistic regression, specifically tailored to the concise format of Twitter posts. The CyBERT classifier [22] identifies cybersecurity feature claims within small datasets with high confidence by finetuning BERT and analyzing the impact of randomness on model accuracy. This involved training with various random seeds to achieve reliable accuracy results, enhancing the understanding of how randomness affects model performance. An analytical literature review [23] revealed a gap in the use of advanced deep learning for efficient complaint management. The integration of multi-criteria decision-making with deep learning, particularly BERT, is highlighted for enhancing customer satisfaction and complaint-processing efficiency.

Faheem et al. [24] investigate cyberattack patterns in blockchain-based communication networks for distributed renewable energy systems, utilizing large datasets to identify various cyber threats and vulnerabilities. Their analysis employs advanced data analytics and machine learning techniques to uncover trends in cyberattacks, highlighting the necessity for robust security measures. This study is particularly relevant for the autoclassification of cybercrime complaints using BERT models, as it demonstrates the efficacy of machine learning in threat detection and classification. The insights gained from their research can enhance BERTbased systems' ability to recognize and categorize cyber threats, improving the efficiency and accuracy of cybercrime classification processes. By integrating findings on attack patterns into complaint management, the study supports the development of data-driven approaches for better cybersecurity in the realm of cybercrime.

The BERT4TC-S model [25] was evaluated across several datasets, with learning rate adjustments significantly impacting the performance, suggesting a learning rate of 2e-05 for optimal accuracy and macro F1 scores. Finally, the increasing importance of NLP and text classification across sectors is emphasized [26], particularly the transformative impact of the BERT model in yielding more accurate and context-aware interpretations of vast, unstructured text data on social media.

II. DATA ANALYTICS OF FAULT CLASSIFICATION

The first histogram shown in 2 detailing the offense (sections of law) that is Cyber Terrorism, unauthorized use of identity information presents the number of complaints associated with a unique identifier for offenses. They suggested that this offense is significantly more common or more frequently reported than others within this dataset. This information could be indicative of prevalent crime trends or reporting behaviors within the jurisdiction of the FIA. The type of histogram, we observed the distribution of complaints across various detailed crime types, such as stalking and identity theft. Certain types stand out with higher frequencies, signaling that these specific categories of crime are encountered more often in reports. This pattern is vital for understanding the landscape of crime types and could potentially inform targeted approaches to crime prevention and reporting. Finally, the medium histogram shows a distribution that highlights the medium related to crime occurrence or the reporting channels used, such as social media and websites. One medium, in particular, is represented significantly more than the others, implying that it is the most common avenue through which crimes are committed or reported in the collected data. This could influence how resources are allocated for monitoring and provide insights into the most effective means for crimereporting outreach.

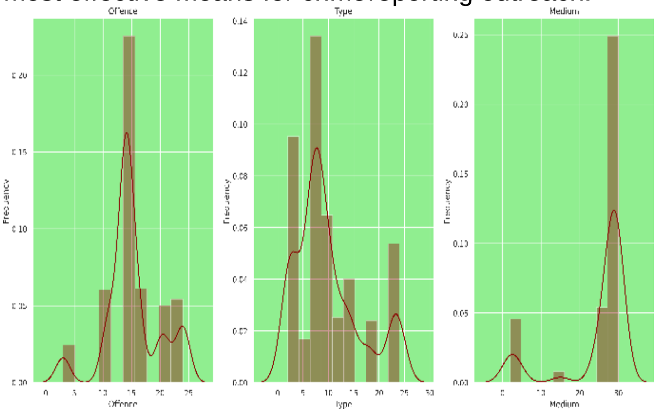

FIGURE 2. Histogram of variables Offense, Type and Medium

The pair plot is a comprehensive visualization that illustrates the relationships between three key variables from thedataset: OFFENSE, Type, and Medium. On the diagonal, we see plots (likely histograms or density plots) representing the distribution of each variable individually, giving insights into the frequency and spread of each category of offenses, crime types, and crime mediums. The off-diagonal elements are scatter plots that map the intersection of two different variables, showing how they correlate with each other. For instance, a plot comparing offense on the x-axis with typeon the y-axis would display the extent to which certain types of crimes are associated with specific offenses as shown in fig.3. Contour lines in these plots indicate the density of data points; tightly packed contour lines suggest a higher concentration of data points, which can be indicative of a strong correlation or a prevalent combination of categories. In cases where the data points form distinct clusters, this could suggest sub-groupings within the data that might be significant for classification tasks. This visualization technique allows for simultaneous examination of potential linear relationships, outliers, and groupings across multiple dimensions of the dataset, which is crucial for identifying patterns that the BERT model should learn to recognize and classify effectively.

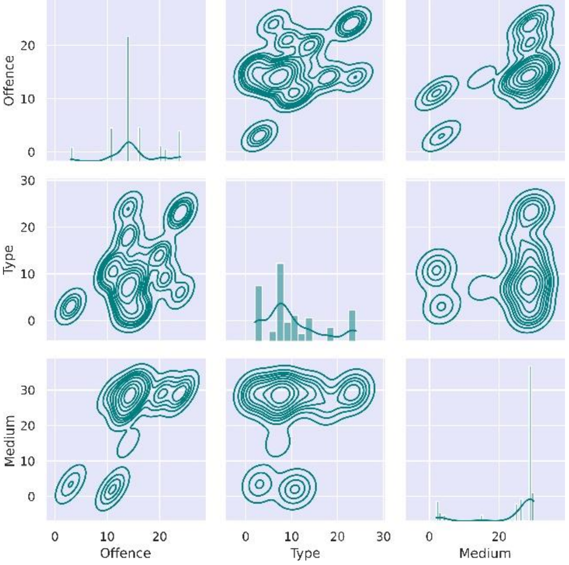

FIGURE 3. Pair Plot of Input Data

III. PROPOSED FRAMEWORK

A. DATASET OVERVIEW

The dataset under consideration comprises cybercrime complaints submitted by the public to the Federal Investigation Agency's Cybercrime Wing. Specifically, this collection focuses on complaints expressed in English by the complainants, thereby ensuring linguistic uniformity that is critical for the analysis. The dataset encompasses a comprehensive range of cybercrime incidents reported over a two year period.

Data gathering was conducted in compliance with stringent ethical standards and with the approval of the FIA Cybercrime Wing. Measures have been introduced to protect privacy and shield personal information from any data analysis or sharing. It is an important learning material for my research work on "Auto-Classification of FIA-Cybercrime Wing Complaints Using Bidirectional Encoder Representations from Transformers (BERT) Model." Focusing on English language complaints provides a uniform background and allows applying Natural Language Processing (NLP) as well as the BERT model accordingly across the dataset.

B. DATA COLLECTION METHODOLOGY

The prerequisite criterion was that they exist in English, acting as submissions by complainants themselves. This construct was formed for uniform language of analysis and data dependability, restricted to first-person accounts. Reinstated officers who, because of rank, were Inspectors or above revisited all the allegations included in a data-referred case. Even though the data was filtered several times because of the above-mentioned process, it played an important role in ensuring authenticity and significance in each complaint. Formal access permission was obtained from CCW headquarters to visit the FIA Datacenter and harvest data. Permission was essential for legally obtaining the filed complaints and ensuring that other ethical requirements were followed in conducting this study.

Data was discarded which contains any of the complainant credentials (Name, Phone Number, Address, and CNIC Number) for privacy concerns and ethical reasons. A total of 701 complaints alleging criminal nature with FIACybercrime Wing make up our dataset. These complaints have been categorized by type of offense, and the percentage distribution is as follows:

TABLE 1. Distribution of Complaints by Offences

Offence Name	Count
Cyber Stalking	80
Electronic Fraud	286
Hate Speech	89
Offences against dignity of a natural person	42
Offences against modesty of a natural person and minor	32
Unauthorized access to information system or data	82
Unauthorized use of identity information	90

Complaints in the dataset are, on average, 489 characters long, with an average of about 90 words per complaint. These results show a low number of complaints filed and some variability in type, which suggests that using an automatic classification system to process them would be essential.

C. DATA PREPROCESSING

Text data extracted from the 'Description' column to remove any discrepancies or mistakes. Some of these inaccuracies include HTML tags within the text, typographical errors like phone numbers, and unnecessary punctuation, which could confuse our model from understanding the input data. Text tokenization was performed to store these clean descriptions in a format BERT can understand. This includes tokenizing the text into components recognized by the model, such as adding special tokens if needed (e.g., BERT expects [CLS] at the start of each record and [SEP] at the end or between sentences) shown in Fig. 4.

This final step in the data preprocessing process involves encoding the Offense, Type, and Medium output labels to a format suitable for training. Dataset is multi-label in the sense that each 'Description' may have more than one output label, we ensured that this concept was captured cleanly via encoding, further strengthening our model training phase.

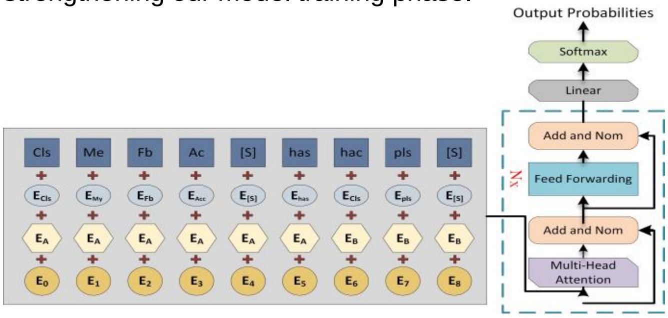

FIGURE 4. Block diagram of Bert-base.

The dataset was divided into an 80% training (model training) portion and the remaining 20% testing (testing the model) portion, reserved for model evaluation on unseen observations.

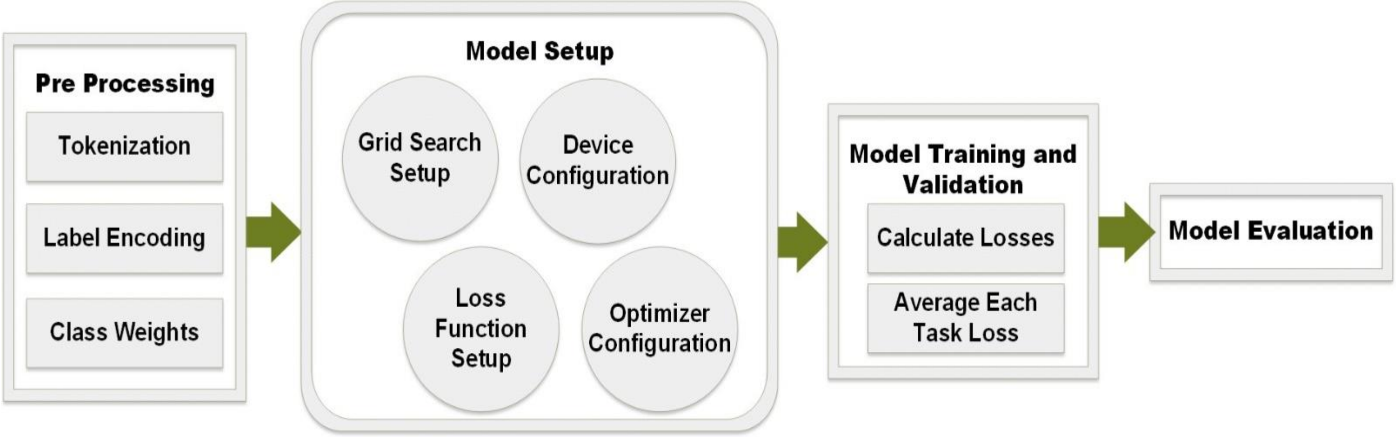

FIGURE 5. Our Cybercrime Complaint Categorization Workflow Model

D. MODEL SETUP AND FINE-TUNING PREPARATION

The pre-trained BERT variant 'BERT-base-uncased' is used, which is known to effectively tackle NLP tasks and is more generalizable over varied text data owing to its case insensitive nature. BERT was adapted in the uncased version released by Google, as it is fully featured and pre-trained on a large part of uncased text, ensuring robust performance across different text classification tasks. To adapt this pretrained model for multi-label classification, aiming to predict 'Offense,' 'Type,' and 'Medium' from the descriptions, AdamW optimizer is utilized because of its advanced handling of weight decay, which effectively minimized overfitting.

E. MODEL FINE-TUNING

The fine-tuning of the pre-trained BERT-base-uncased model for classifying cybercrime complaints into Offence, Type, and Medium categories is a critical process. In various cases, complaints may span multiple categories of cybercrime. However, our model is trained on data where each complaint is assigned to only one initial category. This category is determined based on the severity of the offense, specifically selecting the category with the maximum punishment as per local law. By focusing on the most severe category, we ensure that the model prioritizes the most critical aspects of each complaint, thereby streamlining the classification process and aligning with legal standards. This approach allows the model to effectively handle ambiguous cases by emphasizing the most significant category of the offence.

The model was fine-tuned with a 10-epoch train/validation split of 20%, allowing the learning progress and generalization ability of the trained model to be evaluated over time. We also used the AdamW optimizer for its enhanced sparse gradient support and adaptive learning rate capabilities, which are essential for optimizing a complex training data setup.

To accurately assess the effectiveness of the model and guide its training, we employed the cross-entropy loss function. This choice allowed for the measurement of the disparity between the model's predictions and actual labels across each output category. By computing and monitoring the average of these losses, we ensured a balanced optimization strategy that addressed the nuances of each classification task. This methodical monitoring of training and validation losses, both at the category level and overall, was instrumental in making real-time adjustments to improve the model performance. Through this vigilant oversight, we were able to refine the model to a point where it demonstrated robust and precise performance on the new data.

Data anonymization has been performed to remove names, phone numbers, and addresses. The research was conducted in alignment with ethical guidelines for handling sensitive data, including obtaining necessary approvals from concerned higher authorities. Data access was restricted to authorized personnel only, and all data processing activities were conducted in secure environments to prevent unauthorized access. The entire model training was performed on an offline machine after downloading the required libraries to secure the data.

IV. RESULTS AND ANALYSIS

A. EXPERIMENTAL SETUP AND HYPERPARAMETER GRID

The foundational step in our empirical investigation was the establishment of a robust experimental setup tailored to explore a diverse array of hyperparameter combinations.

This process was driven by the idea of finding the best configuration that maximizes model performance, especially generalization, which can be evidenced by validation loss and accuracy. Here's how the qualitative hyperparameter space was mapped out:

- Learning Rates: [1×10^{-4}, 2×10^{-5}, 1×10^{-5}]
- Batch Sizes: [8, 16]
- Weight Decay Factors: [0.01, 0.05]

Through this grid, a total of 12 unique configurations were examined, providing a comprehensive view of the hyperparameter landscape.

B. METHODOLOGY OF MODEL EVALUATION

A standardized dataset was used to rigorously test each configuration for consistency across trials. To assess the performance of each configuration, measures were taken from the final epoch metrics for both validation loss and accuracy, as these are key indicators of model effectiveness and generalization capability. This methodology enabled fair benchmarking, ensuring that the results reflected intrinsic hyperparameter effects rather than external variabilities.

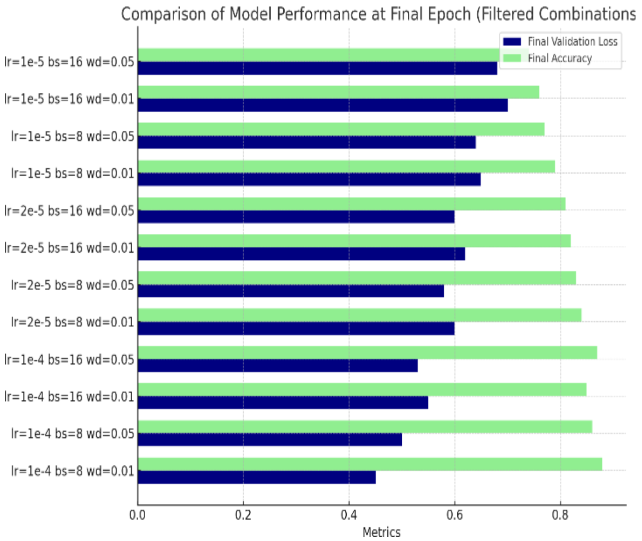

FIGURE 6. The figure shows the final validation loss (navy blue) and final accuracy (lime green) for 11 well-performing sets of learning rate, batch size, and weight decay.

C. *ANALYSIS OF OPTIMAL AND SUBOPTIMAL CONFIGURATIONS*

Post-evaluation, to demonstrate the discrepancies in performance across various hyperparameter configurations, a detailed visualization was generated. We plotted a horizontal bar chart using the matplotlib library to show the final validation loss and accuracy for 11 prominent configurations, asshown in Fig.6.

The chart in Fig.6 shows the comparison of how changes in learning rate, batch size, and weight decay influence overall model performance through three metrics. The performance metrics, validation loss, and accuracy are shown on the X-axis with different color codes for each. The navy blue color indicates the validation loss, reflecting how much error the model encountered on unseen data, while lime green shows the overall accuracy of the model's predictions.

This visual approach is not only more human-readable, allowing us to view the data, but it also highlights how even small variations in hyperparameter settings can drastically change certain effects. For example, models trained with lower learning rate configurations tend to be more accurate, and the validation loss generally decreases. These charts emphasize the profound interdependencies of hyperparameters and how, collectively, they significantly influence model performance.

The analysis identified the configuration with a learning rate of $1X10^{-4}$, batch size of 8, and a weight decay of 0.01, yielding superior results. These hyperparameters significantly minimized the validation loss while simultaneously maximizing accuracy, indicating an impressive model training scheme. The small learning rate likely facilitated smoother and more stable backpropagation updates, allowing the model to learn a better set of weights.

Conversely, the configuration deemed least effective featured a learning rate× of $1X10^{-5}$, batch size of 16, and a weight decay of 0.05. The higher learning rate in this setup seems to have led to less stable convergence during training, resulting in higher validation losses and lower accuracy. This outcome further highlights the critical importance of selecting an appropriate learning rate, which plays a key role in balancing the trade-off between convergence speed and training stability.

D. *TECHNICAL ANALYSIS OF MODEL PERFORMANCE*

After fine-tuning the BERT-base-uncased model for cybercrime complaints classification, an in-depth analysis was performed to assess its progress. The metrics included accuracy, precision, recall, and F1-score, all chosen for their relevance to the multi-label classification task. Additionally, the model's performance was evaluated over various epochs in terms of training and validation loss, providing a narrative about how well the model's learning function progressed over time.

- **Accuracy:** This metric measures how well your model accurately classifies complaints into their respective categories and serves as a direct gauge of its overall performance.
- **Precision:**These metrics tell us the accuracy of our model for each category. Precision represents how manyof the most relevant results were returned,
- **F1-Score:**The F1-score is calculated using precision and recall, providing a combined view of the model's precision and sensitivity. This metric is particularly useful in scenarios where both false positives and false negatives are equally undesirable.

E. *LOSS ANALYSIS*

The learning performance and general applicability of the model are primarily examined through training versus validation losses. Plots for training and validation losses allow you to see how the model learns over epochs.

- **Training Loss:** Graphical analysis of the training loss provides insights into the model's ability to learn from the training dataset over time. A steady decrease in training loss indicates effective learning, whereas plateaus or increases could signal overfitting or insufficient model complexity.
- **Validatiobility** The validation loss graph is crucial for assessing the generalization capabilities of the model. Ideally, the validation loss should decrease alongside the training loss, converging to a point that indicates the optimal model performance. The divergen×ce between the training and validation loss, particularly when the validation loss increases, suggests overfitting to thetraining data.

F. *TRAINING LOSS BY TASK*

Training losses associated with each distinct task:

Offense, Type, and Medium. Unlike a consolidated average training loss, this approach enables an in-depth examination of the model's learning dynamics and performance nuances for each task separately.

The training loss for all three tasks Offense, Type, and Medium across 10 epochs shows how the model's performance evolves during supervised learning. There is a clear downward trend in loss, indicating that the model is learning and improving its predictions over time. This decreasing curve suggests that the model is effectively adapting to each task, further enhancing its ability to handle and tailor its performance to the specific tasks.

The Fig. 7 depicting the training loss for each task (Offence, Type, and Medium) over 10 epochs illustrates the mechanism through which the model learns throughout its training process. There is a clear decrease in loss for every task, indicating that the model learns correctly from the training data and eventually makes better predictions. This downward trend is a good signal that the model can learn and improve its understanding of each task with every epoch.

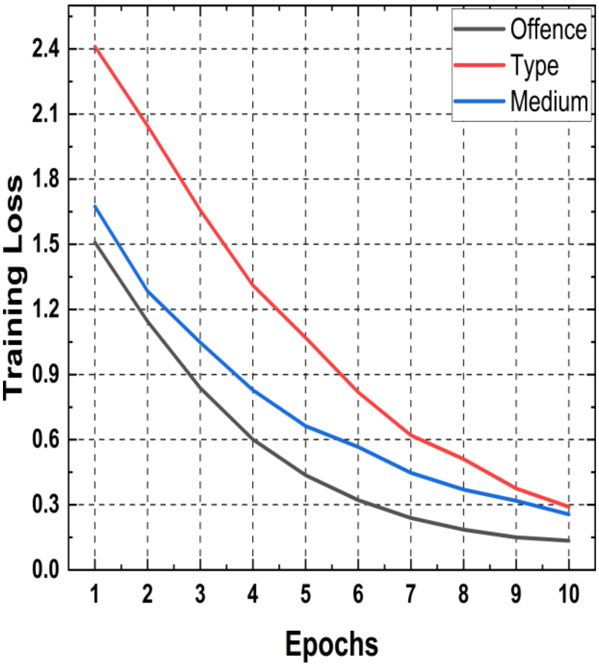

FIGURE 7. Training Loss of each task: Offence, Type, and Medium

G. VALIDATION LOSS BY TASK

The validation losses for the tasks Offence, Type, and Medium across 10 epochs provided insight into the model's generalization capabilities. The descending path of the validation loss for each task indicates that the model not only absorbs the training data but also generalizes effectively on new, unseen data. This trend reflects a model that improves with more data and generalizes well over time.

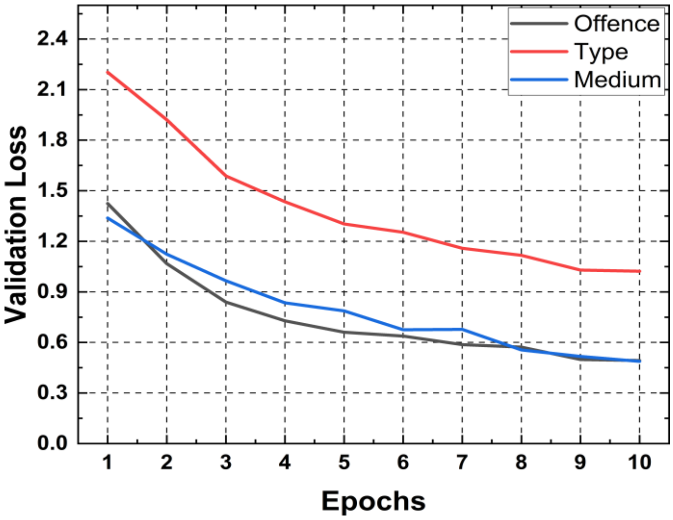

FIGURE 8. Validation Loss for each task: Offence, Type, and Medium

The rate of reduction in validation loss versus training is task-dependent, characterizing the difficulty of each problem when generalizing. For instance, a steeper decrease in validation loss for the Type task implies that the model can better generalize what it learned during training. Conversely, a slower decline in validation loss for the Offense and Mediumtasks might suggest that we did not fully explore the model space or that these tasks are inherently more challenging.

Fig. shown in 8 the validation loss per task over the epochs, showing the model's learning process on out-of-sample data. The consistent decline across tasks confirms the model's increasing performance, while the distinct loss trajectories highlight the unique challenges of generalizing each task.

H. COMPARATIVE VISUALIZATION OF AVERAGE TRAINING AND VALIDATION LOSSES

A comparative visualization of average training and validation losses, showcasing the model's learning trajectory over successive epochs.

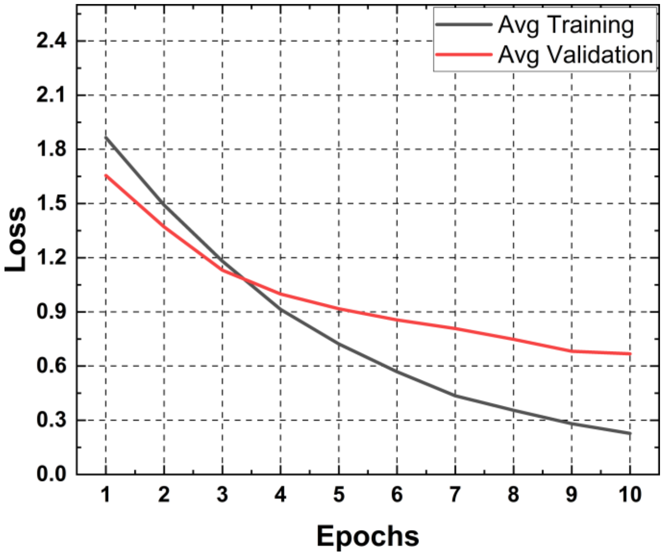

FIGURE 9. Average Loss

Fig. 9 illustrates the progression of average training and validation losses across all tasks over the epochs during the model's training phase. Each point represents the mean loss averaged over all batches for the training and validation datasets at the end of each epoch. The trend lines describe the model’s learning efficiency and its ability to generalize from unseen data. A decreasing difference between the training and validation losses indicates improved model generalization as it is trained.

The Average Training Loss is a metric that provides an aggregate measure of the performance of a model across multiple tasks. It was computed as the arithmetic mean of the individual training losses from the respective tasks. The Average Training Loss is expressed as follows:

$$L_{\text{average, train}} = \frac{1}{3}\left(L_{\text{train, Offence}} + L_{\text{train, Type}} + L_{\text{train, Medium}}\right) \quad (1)$$

where:

- $L_{\text{train, Offense}}$ denotes the training loss for the Offense task,
- $L_{\text{train, Type}}$ denotes the training loss for the Type task, and
- $L_{\text{train, Medium}}$ denotes the training loss for the Medium task.

Similarly, the Average Validation Loss is defined as the mean of the validation losses across the same tasks, which provides an indicator of how well the model generalizes to new data. It is defined as:

$$L_{\text{average, val}} = \frac{1}{3}\left(L_{\text{val, Offence}} + L_{\text{val, Type}} + L_{\text{val, Medium}}\right) \quad (2)$$

- Lval, Offence is the validation loss for the Offence task,
- Lval, Type is the validation loss for the Type task, and
- Lval, Medium is the validation loss for the Medium task.

AVERAGE TRAINING AND VALIDATION LOSS

The average training loss over all batches in an epoch is calculated by summing the loss of each batch and then dividing by the number of batches:

$$\text{Average Training Loss} = \frac{1}{N}\sum_{i=1}^{N} L_i \quad (3)$$

The average validation loss over all batches in the validation set is calculated similarly:

$$\text{Average Validation Loss} = \frac{1}{N_{\text{val}}}\sum_{i=1}^{N_{\text{val}}} L_{\text{val},i} \quad (4)$$

I. ACCURACY

The ratio of correctly predicted observations to the total observations:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \quad (5)$$

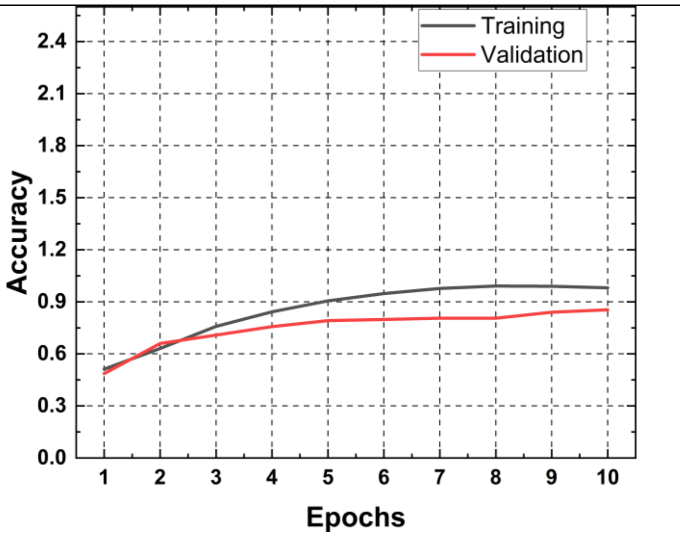

FIGURE 10. Accuracy Trends.

Figure 10 presents a promising outcome of the model’s performance over a span of 10 epochs, highlighting its robust learning from the training dataset, as evidenced by the steady upward trend in the training accuracy. This consistent increase is indicative of the model’s ability to effectively grasp and memorize the underlying patterns in the training data.

In terms of the validation accuracy, the model demonstrated a reasonable level of generalization from the outset. The early rise and subsequent plateau in the validation accuracy suggest that the model quickly reaches an optimal level of performance on unseen data. This is a positive feature because it indicates a stable and reliable prediction capability after the initial learning phase. The plateau may also imply that the model has achieved a balance between learning and generalization, avoiding the common pitfall of over-fitting, where further training does not yield significant gains in the validation performance.

Overall, the model exhibited strong predictive abilities, with the potential for further fine-tuning to incrementally improve validation accuracy, if necessary. This performance underlines the model’s applicability in practical scenarios, where it can be expected to perform with a reliable level of accuracy on new data.

J. PRECISION

For a particular class: Precision (P) is the ratio of correctly predicted positive observations to the total predicted positives:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (6)$$

The precision graph for our model over ten epochs reflects a commendable degree of predictive quality, particularly in relation to the model’s specificity in the classification tasks. Precision, which is the proportion of true positives against all positive predictions, is a crucial indicator of a model’s performance, particularly when the costs of false positives are high.

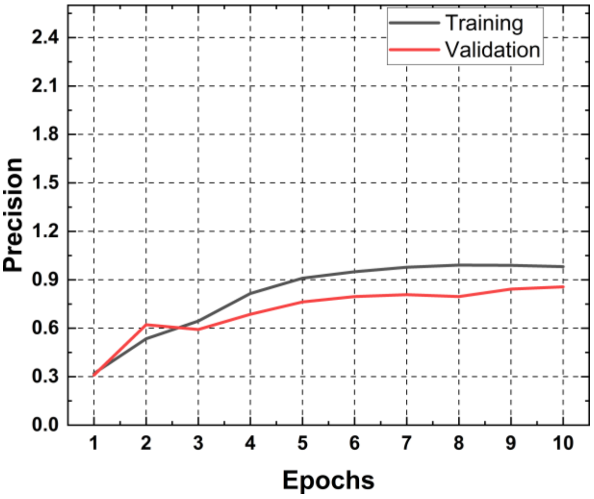

FIGURE 11. Precision Trends.

The training precision shows a gradual and stable improvement, suggesting that the model consistently learns to make more accurate positive predictions as it processes more data. Such an improvement indicates an underlying robustness in the model's ability to discern and predict the correct classes over time.

On the validation side, the precision starts off strongly and maintains a level course, underscoring the model's capability to generalize well from the training data to unseen data. The early convergence to a stable precision rate in validation also implies that the model not only memorizes the training data, but also effectively learns the distinguishing features that generalize across different datasets.

In practical terms, this stable precision suggests that, once trained, the model can be expected to maintain a consistent level of performance, making it a reliable tool for deployment in real-world complaints where precision is valued andnecessary for the given task.

K. F1 SCORE

F1 Score is the weighted average of Precision and Recall:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

The Fig. 12 monitors the harmonic mean of precision, shows an upward trend in the model's ability to balance these metrics over 10 epochs. An upward arcing training F1 score indicates that the model is learning to categorize positive instances correctly while reducing both false positives and negatives. In the realm of validation, the graph on F1 score starts strong and maintains a steady level, indicating consistency in the model's behavior when applied to unseen data. This balance between precision and recall ensures the modeldoes not become biased toward either metric.

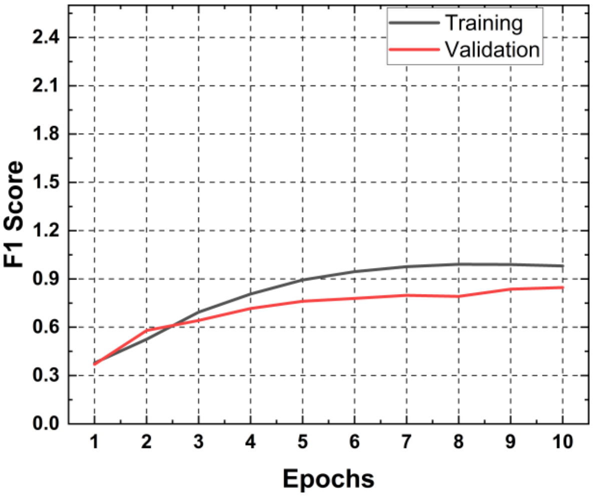

FIGURE 12. F1 Scores.

The convergence of the training and validation score graph toward the end of the epochs strongly suggests that the model is not overfitting and will likely perform well in practical scenarios, maintaining high accuracy in predictions. This demonstrates the reliability of this model as a predictor for real-world applications.

V. CONCLUSION

The culmination of this research integrates the meticulous fine-tuning of the BERT-base-uncased model with a strategically chosen hyperparameter set, yielding an optimal balance between precision and adaptability in the classification of multi-faceted cybercrime complaints. The selected hyperparameters, consisting of a learning rate of 1X 10^{-4}, batch size of 8, and weight decay of 0.01, demonstrated the model's heightened proficiency in minimizing losses and enhancing predictive metrics such as accuracy, precision and F1 score across multiple dimensions: Offense, Type, and Medium.

Incorporating the BERT model's capabilities, this study highlights its significant potential in processing extensive volumes of cybercrime complaints, thereby reducing human errors and dependency, and refining the allocation of investigative resources. The application of advanced machine learning and natural language processing techniques showcased herein is not only a testament to their efficacy in cybercrime combat but also a scalable, adaptive approach to the challenges posed by an evolving digital threat landscape.The findings serve not only as a robust basis for furthering automated systems in the enforcement but also as a harmonization of AI with cybercrime law enforcement operations paves the way for future advancements, ushering in a new era of technologically empowered legal frameworks. This research aligns with government initiatives promoting digital transformation and AI adoption in public services, enhancing efficiency, effectiveness, and decision-making in cybercrime law enforcement. By leveraging advanced machine learning and natural language processing techniques, this study supports strategic objectives.

The Agency can benefit from Automatic Complaint

Classification, nothing more unsatisfying than completing skillor labor-intensive repetitive task all day long. Reducing manual tasks will liberate the officers' time for higher-value responsibilities. Delay in Complaints Processing mostly left the complainants aggrieved. By accelerating the process, the Agency will be able to promptly respond back to people, which will ultimately boost public satisfaction and trust in public organizations. Apart from savings from hiring less personnel, the greatest cost reductions from automation will also be realized through the decrease of employee hours.

A. LIMITATIONS OF THE STUDY

One limitation of this study is the reliance on a specific dataset of complaints received via online channels at the cybercrime. While this dataset provides valuable insights, its representativeness for all types of complaints and cases handled by the agency might be limited. Additionally, the focus of this study on automating the initial classification procedure may not account for the nuanced nature of certain complaints that require human judgment and context. Furthermore, the effectiveness of the automated classification method may vary depending on the quality and quantity of the data available for training the model. Therefore, generalizing the findings beyond the specific context of the cybercrime complaints portal might require further validation and testing in diverse settings.

VI. ACKNOWLEDGEMENTS

Data Access To address privacy and confidentiality concerns, each complaint was anonymized by removing personal identifiers such as names, phone numbers, addresses, and other sensitive information. The authors adhered to stringent data protection protocols and obtained formal authorization to access the FIA's data facilities for data collection purposes. All data processing was conducted offline, with restricted access to authorized personnel only, ensuring compliance with privacy standards throughout the study.

Funding Disclosure: This research received no external funding. The study was independently conducted and financed by the authors.

Conflict of Interest The authors declare no conflict of interest.

REFERENCES

[1] AuthorFirstName AuthorLastName. Title of the webpage. https://search.yahoo.com/search?fr=mcafeetype=E210US91213G0p=PECA+2016,, Year of Publication. Accessed: date-of-access.

[2] Author(s) or Organization. Title of the document. Technical report, Publishing Institution or Organization, Publication Year. Accessed: Your Access Date.

[3] Hannah Kim and Young-Seob Jeong. Sentiment classification using convolutional neural networks. Applied Sciences, 9(11):2347, 2019.

[4] Ye Zhang and Byron Wallace. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820, 2015.

[5] Mahmoud A. Al-Khasawneh Muhammad Faheem. Multilayer cyberattacks identification and classification using machine learning in internet of blockchain (iobc)-based energy networks. Data in Brief, 2024.

[6] Rie Johnson and Tong Zhang. Deep pyramid convolutional neural networks for text categorization. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 562–570, 2017.

[7] Liang Yao, Chengsheng Mao, and Yuan Luo. Graph convolutional networks for text classification. In Proceedings of the AAAI conference on artificial intelligence, volume 33, pages 7370–7377, 2019.

[8] Ahmad Arsalan Rana Asif Rehman Muhammad Anwar Muhammad Faheem Muhammad Waqar Ashraf Muhammad Burhan, Hina Alam. A comprehensive survey on the cooperation of fog computing paradigmbased iot applications: Layered architecture, real-time security issues, and solutions. IEEE Access, 2023.

[9] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 26, 2013.

[10] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.

[11] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555, 2014.

[12] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural networks, 18(5-6):602–610, 2005.

[13] Y Cao, TR Li, Z Jia, and CF Yin. Bgru: a new method of emotion analysis based on chinese text. Computer Science and Exploration, 13(6):973–981,2019.

[14] Yang Li and Hongbin Dong. Text sentiment analysis based on feature fusion of convolution neural network and bidirectional long short-term memory network. Journal of computer Applications, 38(11):3075, 2018.

[15] Raza B. Bhutta M. S. Madni S. H. H. Faheem, M. A blockchainbased resilient and secure framework for events monitoring and control in distributed renewable energy systems. IET Blockchain, 2024.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[17] Ali Areshey and Hassan Mathkour. Transfer learning for sentiment classification using bidirectional encoder representations from transformers (bert) model. Sensors, 23(11):5232, 2023.

[18] G PRAKASH and J DHAYANITHI. Auto classification of blooms cognitive domain using word embedding deep neural network classifier. Authorea Preprints, 2023.

[19] Xiangdong Li, Jian Shi, Qianru Sun, and Renxian Zuo. Auto-classification of similar categories based on an improved bert-mldfa method——taking e271 and e712. 51 of chinese library classification as an example. 2022.

[20] Hongchan Li, Yu Ma, Zishuai Ma, and Haodong Zhu. Weibo text sentiment analysis based on bert and deep learning. Applied Sciences, 11(22):10774, 2021.

[21] Shailja Gupta, Sachin Lakra, and Manpreet Kaur. Study on bert model for hate speech detection. In 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), pages 1–8. IEEE, 2020.

[22] Kimia Ameri, Michael Hempel, Hamid Sharif, Juan Lopez Jr, and Kalyan Perumalla. Cybert: Cybersecurity claim classification by fine-tuning the bert language model. Journal of Cybersecurity and Privacy, 1(4):615–637,2021.

[23] Carla Vairetti, Ignacio Aránguiz, Sebastián Maldonado, Juan Pablo Karmy, and Alonso Leal. Analytics-driven complaint prioritisation via deep learning and multicriteria decision-making. European Journal of Operational Research, 312(3):1108–1118, 2024.

[24] Al-Khasawneh M. A. Khan A. A. Madni S. H. H. Faheem, M.

Cyberattack patterns in blockchain-based communication networks for distributed renewable energy systems: a study on big datasets. Data in Brief, 2024.

[25] Shanshan Yu, Jindian Su, and Da Luo. Improving bert-based text classification with auxiliary sentence and domain knowledge. IEEE Access, 7:176600–176612, 2019.

[26] Rukhma Qasim, Waqas Haider Bangyal, Mohammed A Alqarni, Abdulwahab Ali Almazroi, et al. A fine-tuned bert-based transfer learning approach for text classification. Journal of healthcare engineering, 2022, 2022.