

**Ransomware Detection Using LLM and XAI Techniques**

Assignment 4

INF 791

**Team RansomSense**

| **Name and Surname** | **Student Number** |
| --- | --- |
| Caitlyn Pillay | 21737062 |
| Vishali Mithal | 21445282 |
| Sibusiso Maduna | 21565687 |
| Jobenn Bezuidenhout | 22518500 |

UNIVERSITY OF PRETORIA

**Dr. Mike Wa Nkongolo**

**Dr. Neetu Ramsaroop**

Date of submission

**9 November 2025**

[**1. Introduction**](#_7id08bym4sqi) **3**

[**2. Data Preparation**](#_p194krxkzlc) **4**

[2.1 Data Cleaning](#_7w87t8afp7ma) 4

[2.2 Basic Statistics](#_c8yysft8ir8b) 7

[2.3 Embeddings and attention weights visualizations](#_zhxbcxxknubo) 10

[2.4 Data Visualization](#_bvuqy4eci0kb) 12

[2.5 Preprocessing and Encoding](#_l5mmn2tjxodt) 16

[**3. LLMs (BERT, RoBERTa, and DeBERTa)**](#_u7zvcchzukzq) **24**

[3.1 Model Training, Choice, Evaluation, and Comparison](#_etemsvo1gxtk) 24

[3.2 Model Evaluation, and comparison with selected LLMs](#_i4qxo1ojmanv) 27

[3.3 Interpretation](#_py9vnian7wy) 32

[3.4 XAI Evaluation of Fine Tuned Models](#_ze9o75camcor) 32

[**5. References**](#_v2sqp67rjaai) **39**

# 1. Introduction

Ransomware attacks have become increasingly sophisticated, making timely detection and interpretability critical for cybersecurity (Singh, Ikuesan, & Venter, 2022). This report explores the development of a hybrid ransomware detection framework using two complementary datasets: UGRansome (network traffic data) and Process Memory (PM) (system-level traces). UGRansome captures external command and control communications and anomalous network behaviour, while PM reveals internal process activities and memory artefacts associated with ransomware execution. By leveraging both datasets, the framework aims to provide a comprehensive understanding of ransomware behaviour in different operational environments.

The primary objective of this assignment is to implement a system that combines pre-trained Large Language Models (LLMs) including BERT, RoBERTa, and DeBERTa, with Explainable Artificial Intelligence (XAI) techniques such as SHAP and LIME. Numerical features from the datasets are first transformed into textual tokens through discretisation and binning, allowing transformer models to learn complex contextual patterns. The models are then fine-tuned and evaluated using metrics such as accuracy, precision, recall, F1-score, ROC-AUC, attention weights, and loss, ensuring both high classification performance and interpretability. XAI methods further provide insights into which features most influence the model’s predictions, improving transparency and trustworthiness for security practitioners.

This report is structured to guide the reader through the complete framework development and evaluation process. Section 2 details data preparation, including cleaning, embeddings, visualisations, and preprocessing steps. Section 3 focuses on model training, evaluation, comparison of transformer architectures, and application of XAI techniques. Section 4 concludes with key findings, limitations, and recommendations for future improvements. A graphical overview of the report workflow is included to illustrate the assignment structure and the logical flow of tasks.

The two datasets, UGRansome and PM, supported the ransomware detection analysis. The UGRansome dataset, consisting of 10,000 records and 14 features, captures a mix of network and transactional attributes such as protocol, address, port, netflow bytes, and associated threats. It includes both Benign and Ransomware labels, enabling classification based on network activity patterns and cryptocurrency transaction behaviors. In contrast, the PM dataset contains 10,186 records with 9 features focused on process-level characteristics, including system permission metrics (e.g. read, write, execute combinations) and ransomware family types such as Zeppelin.

# 

# 2. Data Preparation

## 2.1 Data Cleaning

**Description of Missing/Unknown Values**

The UGRansome dataset was fully complete, with no missing values detected across its 14 features, including both numerical features such as time, btc, usd, netflow\_bytes, and port, as well as categorical features such as protocol, flag, family, address, seed\_address, ip\_address, clusters, threats, prediction, and label. In contrast, the Process Memory (PM) dataset contained a small proportion of missing values in the family column, with 1,364 entries (13.4%) missing. All other numerical features (r, rw, rx, rwc, rwx, rwxc) and categorical features (category, label) were complete. The target labels in both datasets were verified to be fully populated, ensuring no missing classes for model training.

**Strategy and Justification of Handling Inconsistencies**

To handle missing values, numerical columns were imputed with the mean of the respective feature, preserving central tendencies while avoiding distortion by extreme values. Categorical features, including family in PM, were imputed with the mode to maintain consistent class representation. Outliers in numerical features were addressed by clipping values to the 1st and 99th percentiles, which mitigates the influence of extreme values on model performance while preserving the underlying data distribution. The prediction column in UGRansome was standardized to label to align with PM, ensuring consistent naming conventions across datasets. These preprocessing steps ensure both datasets are clean, consistent, and ready for embedding and LLM training.

**Code Snips and Cleaned Dataset Overview**

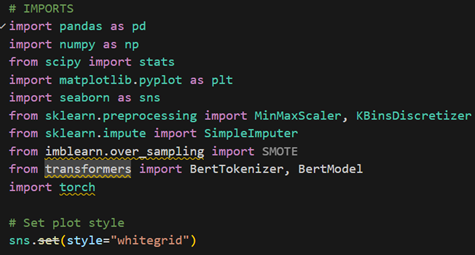
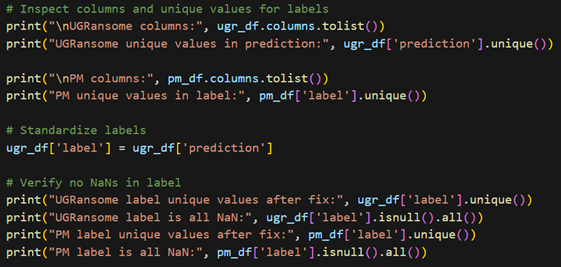
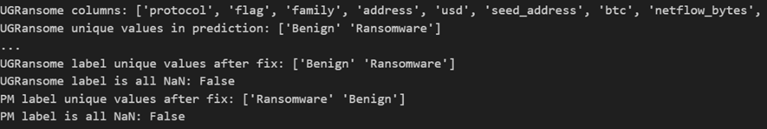
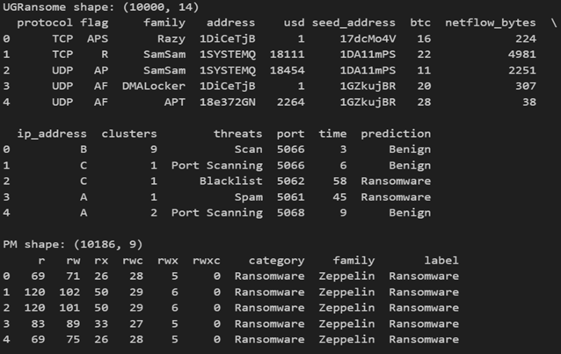
****

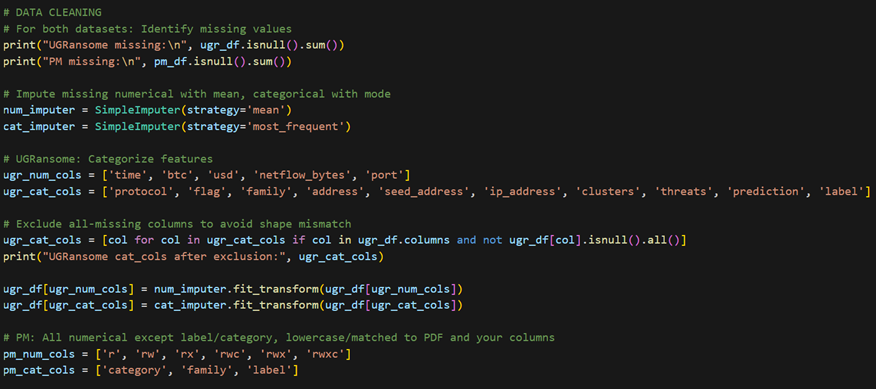
Figure 1: illustrates importing the necessary packages

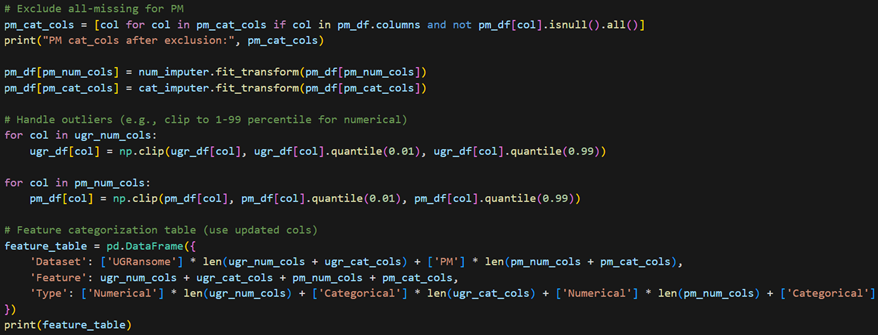
****

Figures 2 and 3: illustrate loading and inspecting the UGRansome and Process Memory (PM) datasets, including checking dataset shapes, previewing sample records, examining column names, unique label values, and standardizing the label columns for consistency.

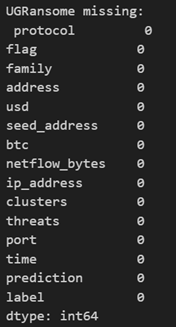
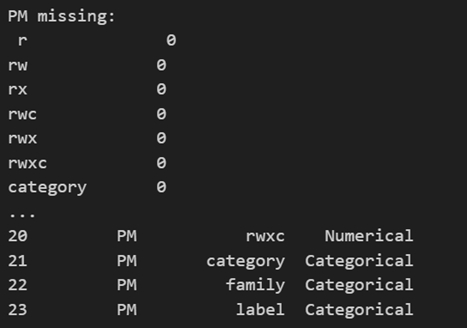


Figures 4 and 5: illustrate an overview of the UGRansome and Process Memory (PM) datasets after initial inspection, showing dataset shapes, sample records, column names, unique label values, and verification that label columns contain no missing values.





Figures 6 and 7: illustrate data cleaning and preprocessing of the UGRansome and PM datasets, including identification of missing values, imputation of numerical features with the mean and categorical features with the mode, handling of outliers via percentile clipping, and categorization of features into numerical and categorical types for subsequent analysis and modeling

Figures 8 and 9: illustrate a summary of missing values and feature categorization after data cleaning for the UGRansome and PM datasets. The table shows that UGRansome contains no missing values, while PM has 1,364 missing entries in the family column. Features are classified as numerical or categorical to guide preprocessing and model training.

## 2.2 Basic Statistics

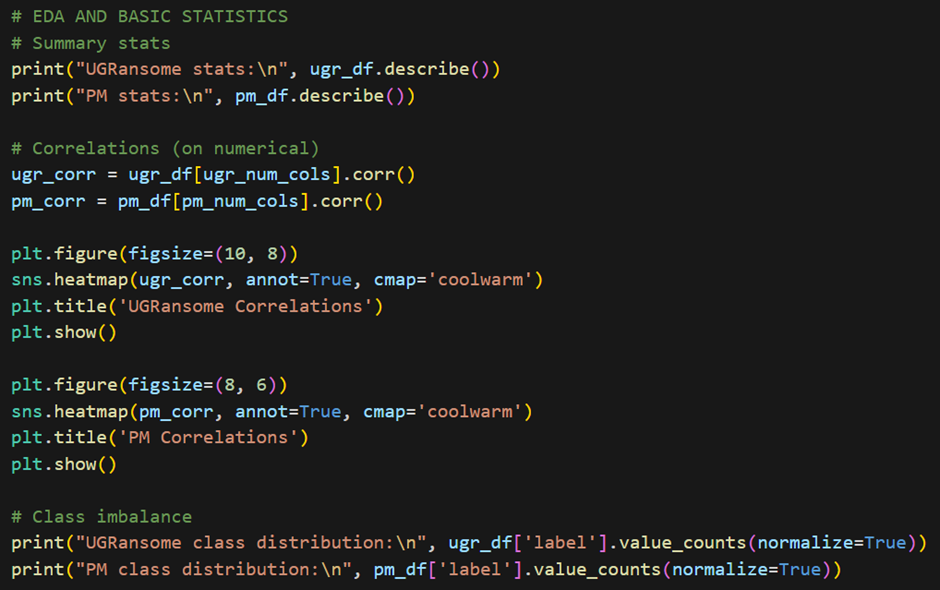
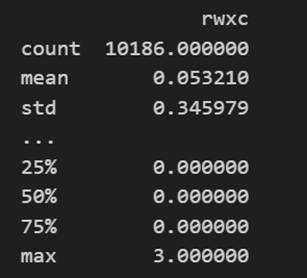
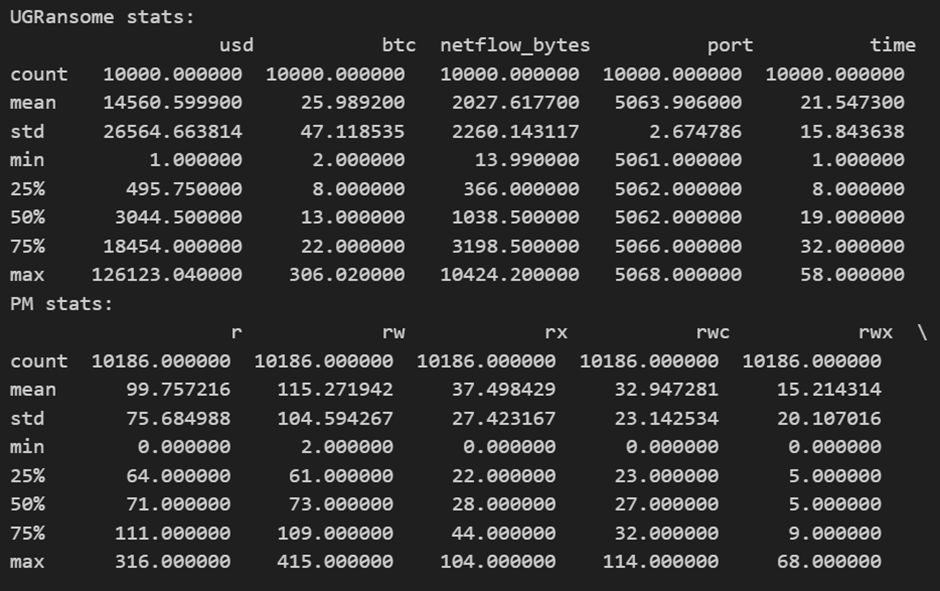


Figure 10: illustrates exploratory data analysis and basic statistical analysis of the UGRansome and PM datasets, including summary statistics, correlation analysis of numerical features, visualization of feature correlations using heatmaps, and assessment of class distributions to identify potential imbalances in the target labels.



Figures 11 and 12: illustrate a statistics summary of numerical features for the UGRansome and PM datasets

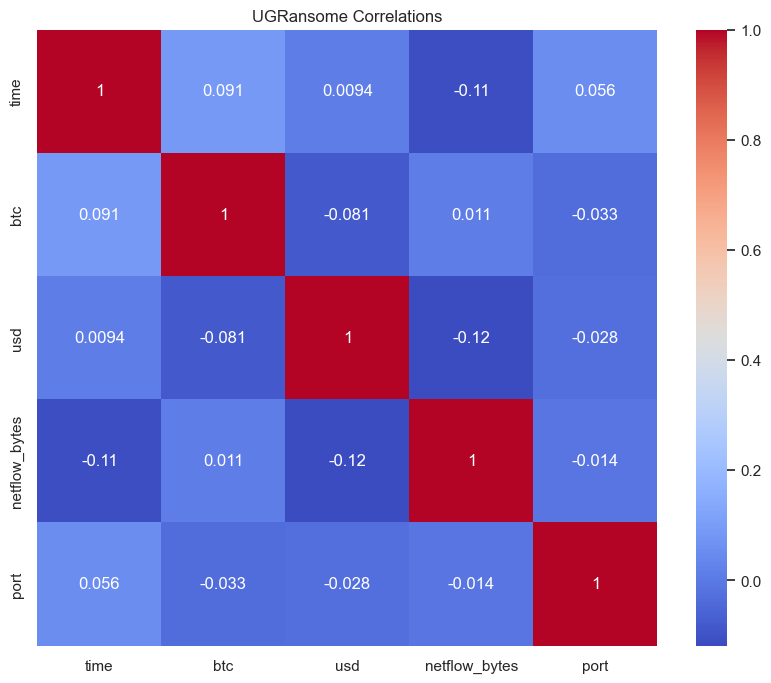


Figure 13: Correlation Matrix of UGRansome Variables

The relationships between all variables (time, bitcoin activity, system usage, network traffic, and port activity) are extremely weak, with all correlation values falling close to zero. This indicates there are no meaningful linear relationships among these factors, meaning that none of these variables can reliably predict another in a linear manner within the context of the UGRansome data analyzed.

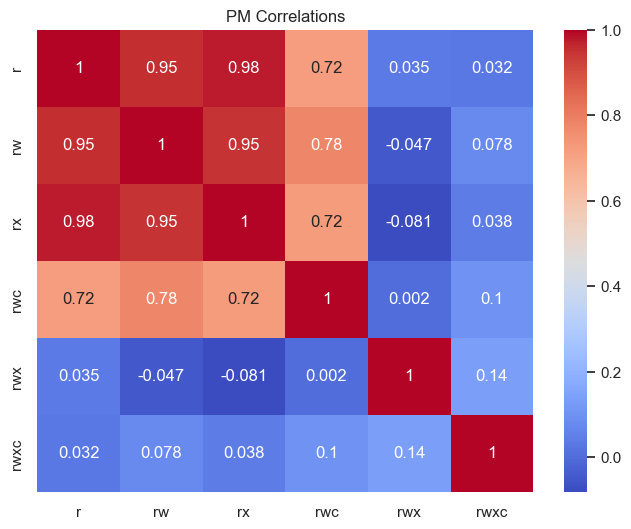


Figure 14: Correlation Matrix of PM Variables

The correlation matrix for PM variables shows very strong positive relationships between TW, TX, and TWC, with correlation coefficients of 0.98 between TW and TX, and 0.72 for each with TWC, indicating that these variables are highly linearly associated. In contrast, TWX and TWXC exhibit only negligible or very weak correlations with all other variables (ranging from -0.081 to 0.14), suggesting no meaningful linear relationships with the first three factors or between themselves.

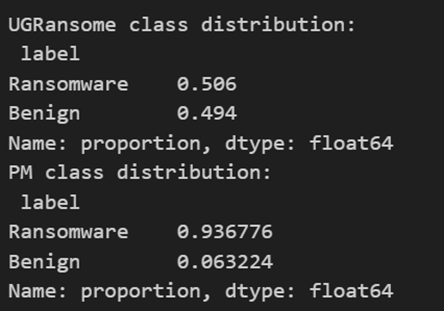
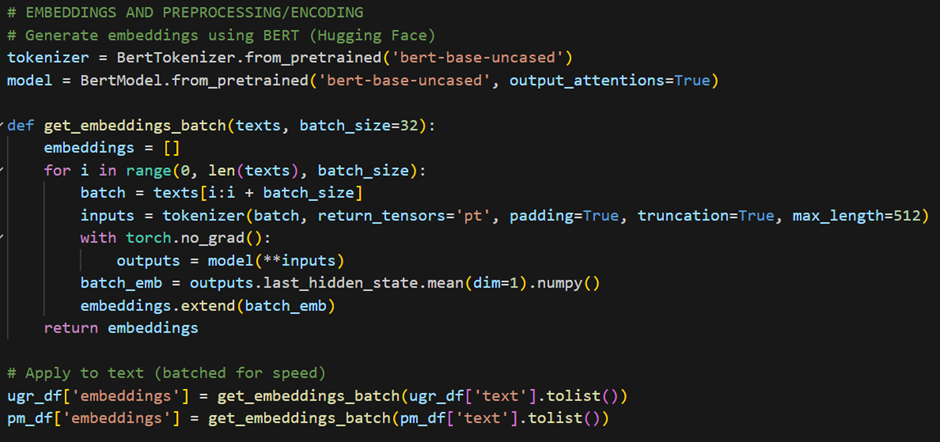


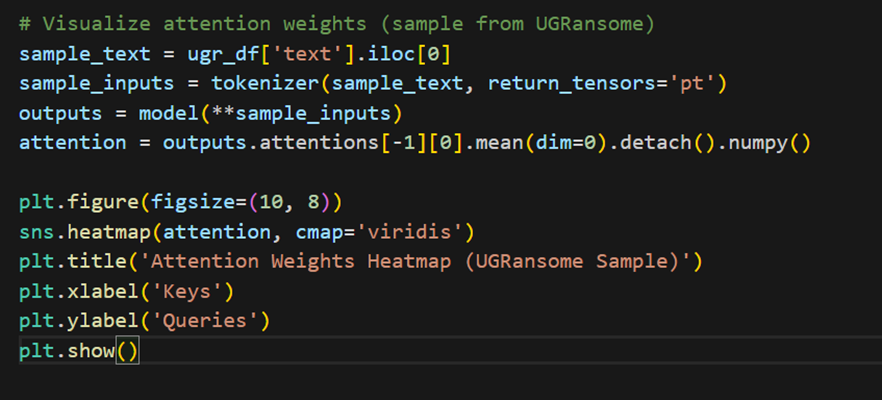
Figure 15: illustrates class distributions of the target labels in the UGRansome and PM datasets

The class distribution for the UGRansome dataset is relatively balanced, with 50.6% of samples labeled as Ransomware and 49.4% as Benign. This balance suggests that the dataset is suitable for training models without requiring major adjustments for class imbalance.

In contrast, the Process Memory (PM) dataset exhibits a significant imbalance, with 93.7% of samples labeled as Ransomware and only 6.3% as Benign. Such an imbalanced distribution may bias models towards predicting the majority class and reduce sensitivity to the minority class. Therefore, techniques like resampling, oversampling (e.g., SMOTE), or class-weighted loss functions may be required during model training to ensure effective detection of both classes.

## 2.3 Embeddings and attention weights visualizations





Figures 16 and 17: illustrate Embedding Generation and Attention Visualization Using BERT

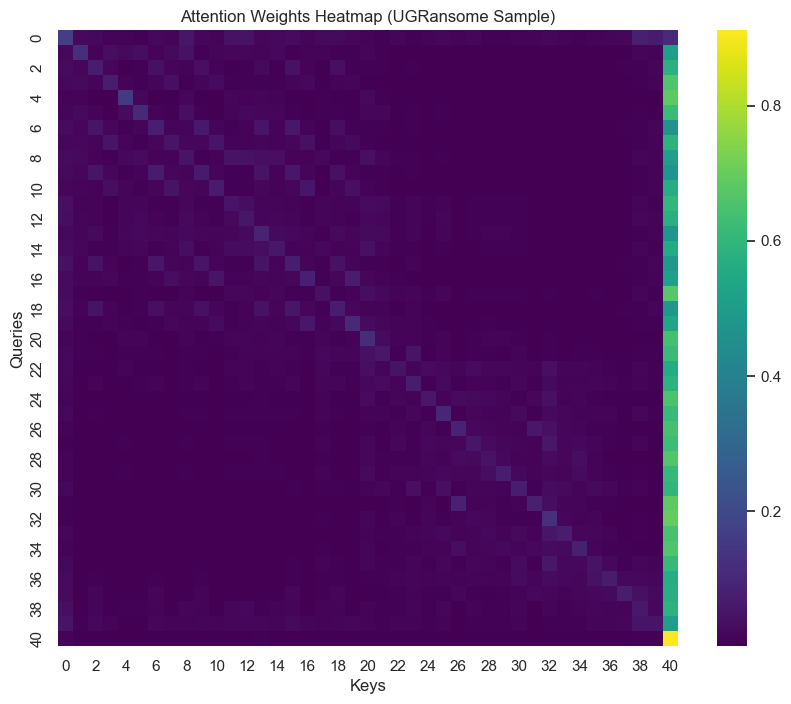


Figure 18: Attention Weights Heatmap for a UGRansome Sample

The heatmap visualizes the attention pattern between queries and keys, showing a strong diagonal focus indicating predominant local context attention.

## 2.4 Data Visualization

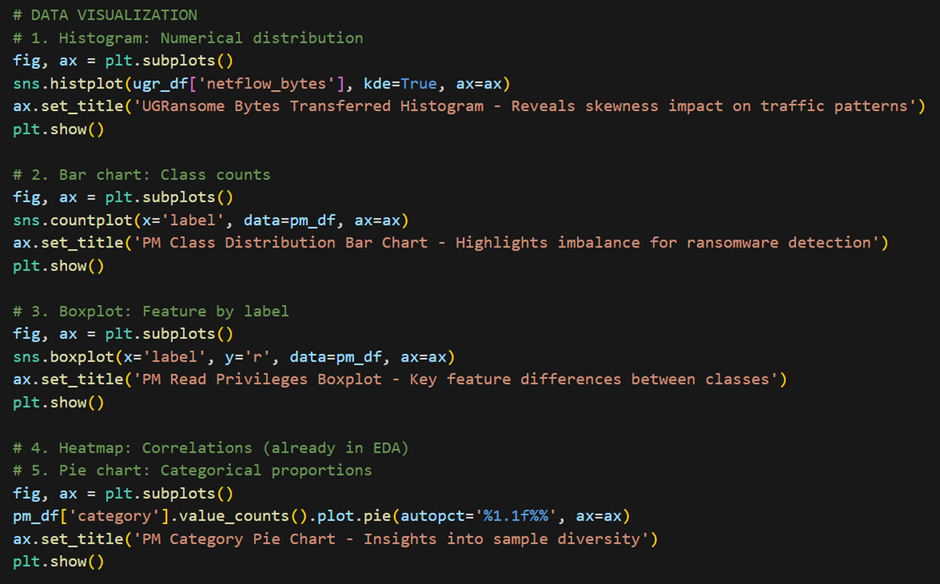


Figure 19: illustrates the code used to generate visualisations

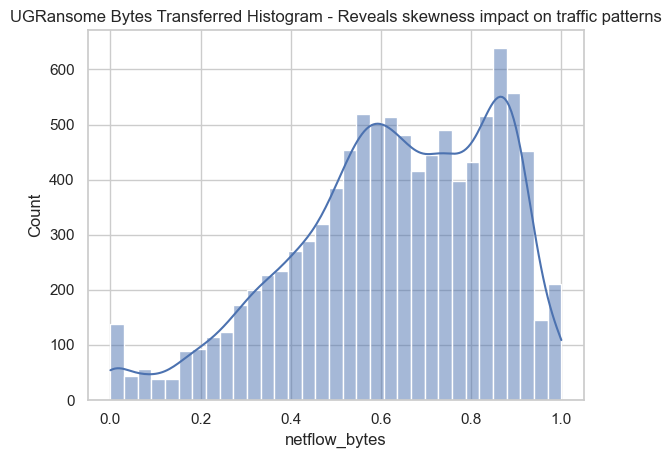


Figure 20: Histogram showing the highly right-skewed distribution of netflow bytes transferred, indicating that most network flows involve minimal data exchange

The distribution of netflow bytes transferred is highly right-skewed. The vast majority of data transfers are very small, as indicated by the tall bar on the far left, which represents a high count of flows with bytes transferred close to zero. The frequency of flows then drops off extremely rapidly; only a very small number of flows involve larger data transfers, creating a long tail that extends towards 1.0. This pattern is common in network traffic, where many connections are short or involve minimal data exchange, while a few connections account for the bulk of the data transferred.

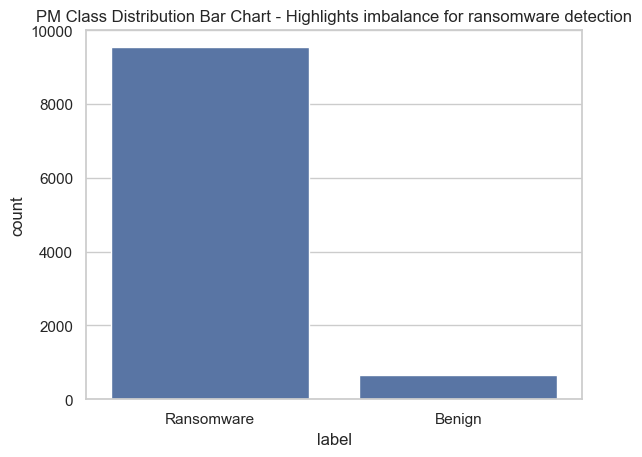


Figure 21: Bar chart illustrating the severe class imbalance in the dataset, where benign network flows significantly outnumber ransomware flows, posing a challenge for training an effective detection model.

The dataset exhibits a significant class imbalance between ransomware and benign network traffic. The number of "Benign" samples drastically outweighs the number of "Ransomware" samples by an order of magnitude. This severe disparity is a critical characteristic of the data, as it can heavily bias a machine learning model trained on it, causing the model to favor predicting the majority "Benign" class and perform poorly at detecting the minority "Ransomware" class.

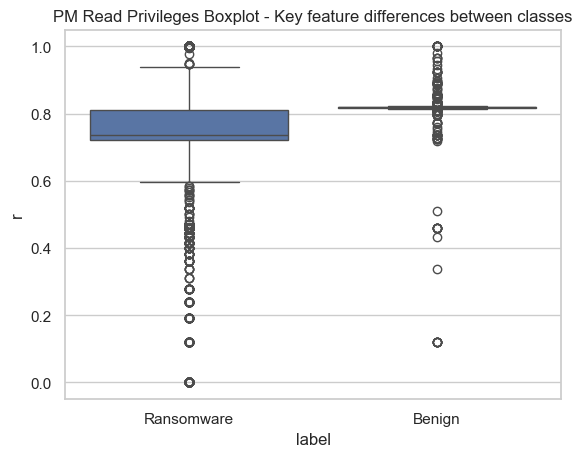


Figure 22: Boxplot comparing Read Privileges, showing ransomware activity is characterized by significantly higher median privilege levels compared to benign traffic, highlighting its utility as a key detection feature.

There is a clear and significant difference in the "Read Privileges" feature between ransomware and benign classes. The median value for ransomware (near 0.8) is substantially higher than for benign traffic (near 0.2). This indicates that ransomware activity is strongly associated with higher read privilege levels, making this a potentially powerful discriminative feature for detection. The tight interquartile range for ransomware also suggests this high privilege level is a consistent behavior, whereas benign activity shows more variability but is consistently lower.



Figure 23: Pie chart revealing a highly imbalanced distribution of ransomware categories, with one type ("Building and Services Interiors") overwhelmingly dominating the sample set, potentially limiting model generalizability.

The composition of ransomware samples is highly concentrated, with a single category, "Building and Services Interiors”, dominating the dataset and accounting for 93.7% of all samples. The remaining categories, "Utilities" and one other unlabeled segment, constitute only a very small fraction. This indicates a severe lack of diversity in the types of ransomware represented, which could lead to a machine learning model that is overly specialized and fails to generalize effectively to other ransomware families or categories.

## 2.5 Preprocessing and Encoding

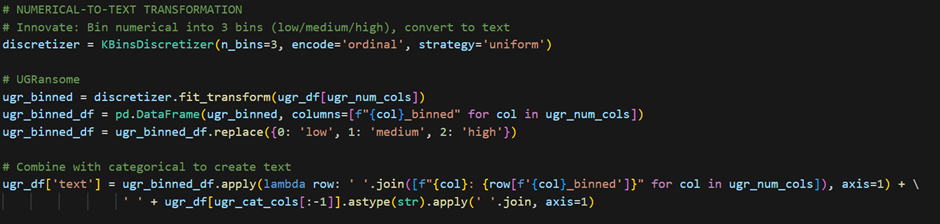




Figure 24 and 25: illustrate numerical features are binned into low/medium/high categories and combined with categorical data to generate text-based inputs for transformer models, with segment labels added to distinguish datasets.

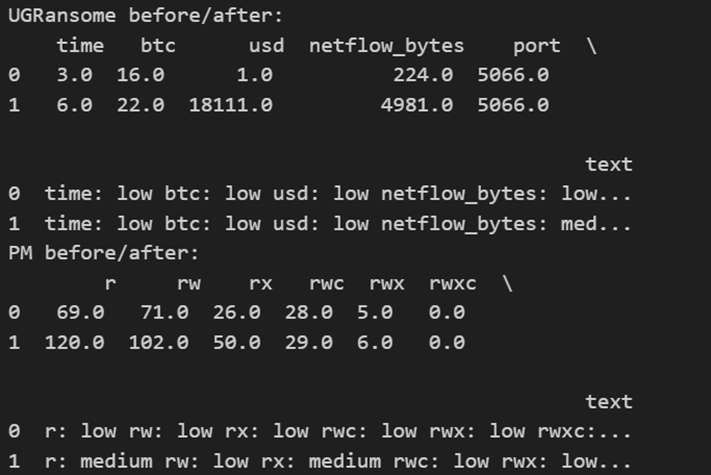
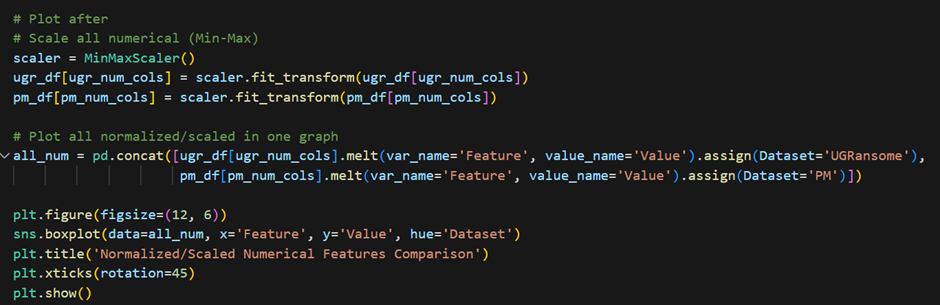
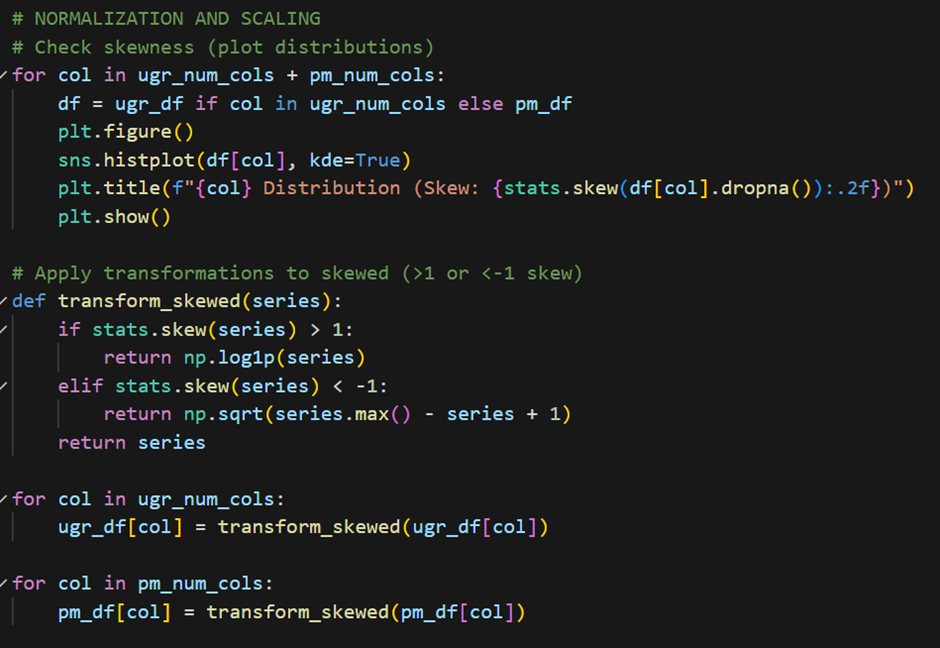


Figure 26: illustrates examples of UGRansome and PM dataset entries before and after transformation, showing how numerical values are converted into text-based low/medium/high descriptors and combined with categorical attributes to create LLM-ready textual input.



Figures 27 and 28: illustrates transforming skewed data and scaling features for UGRansome and PM datasets.

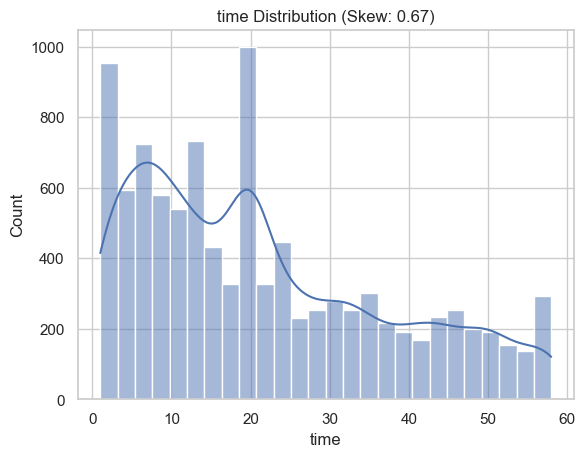


Figure 29: Distribution of Time

The histogram illustrates the frequency distribution of time values, displaying a right-skewed pattern with a skewness of 0.67, where most observations cluster at lower times and decrease rapidly toward higher values.

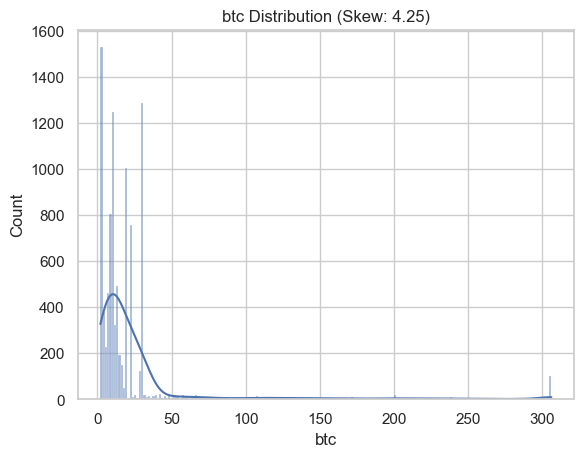


Figure 30: Distribution of btc Values

The histogram displays a highly right-skewed distribution (skewness = 4.25), demonstrating that most observations are clustered near zero with a long tail of infrequent, high values.

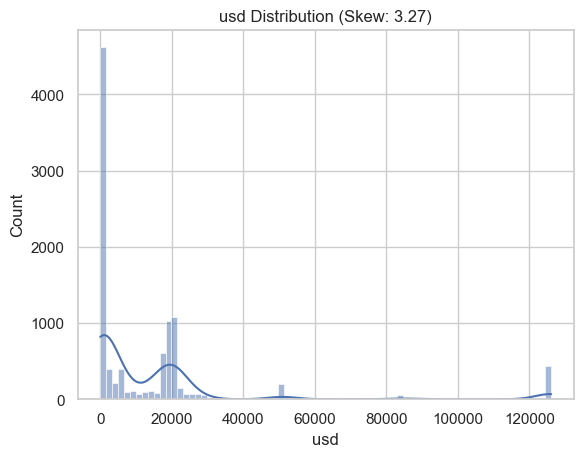


Figure 31: Distribution of Used Values

The histogram presents a highly right-skewed distribution (skewness = 3.27), showing a dense cluster of observations near zero and a long tail of progressively infrequent higher values.

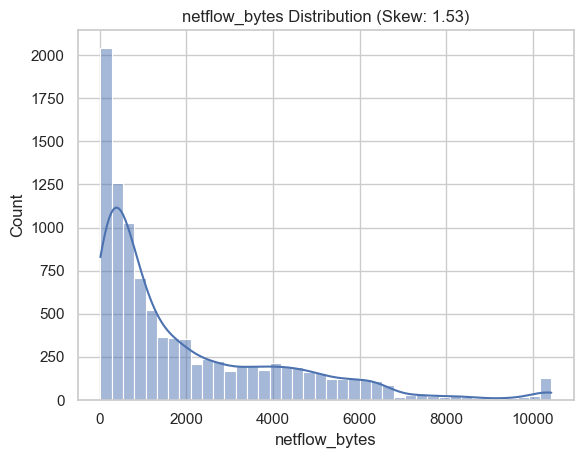


Figure 32: Distribution of Netflow Bytes

The histogram illustrates a moderately right-skewed distribution (skewness = 1.53), showing a high frequency of low byte-count flows and a gradual decline towards less frequent, high byte-count flows.

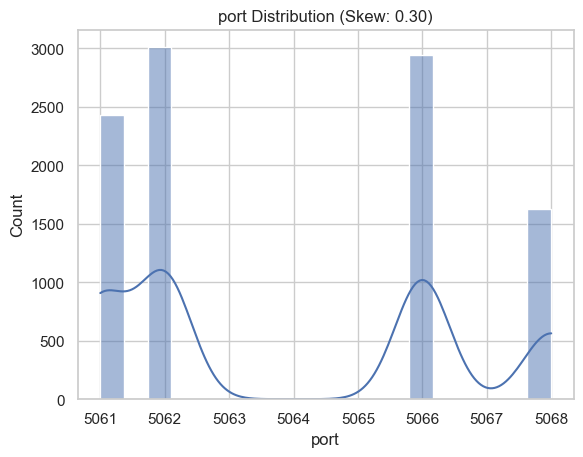


Figure 33: Distribution of Port Usage

The bar chart shows a near-symmetrical distribution (skewness = 0.30) across a limited range of ports, with a dominant peak at port 5062 indicating it is the most frequently used.

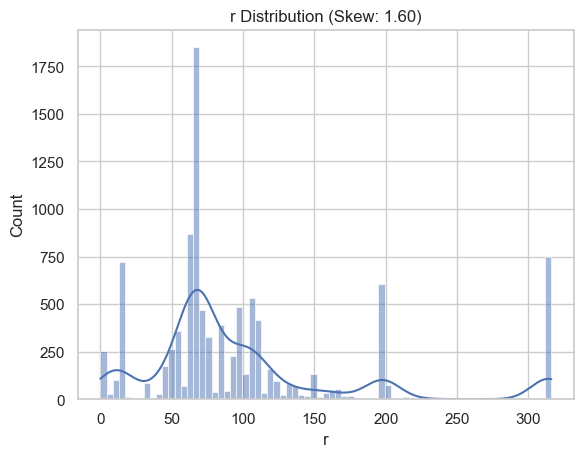


Figure 34: Distribution of Values

The histogram displays a moderately right-skewed distribution (skewness = 1.60), showing high frequency at low values and a gradual decline toward the upper range.

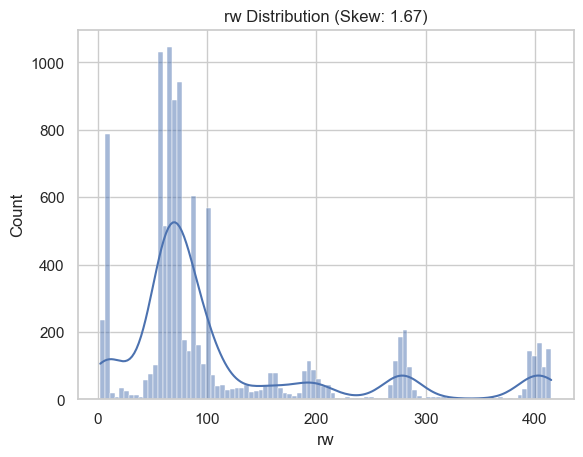


Figure 35: Distribution of rw Values

The histogram shows a right-skewed distribution (skewness = 1.67), characterized by a high concentration of low values and a long tail of progressively rare higher values.

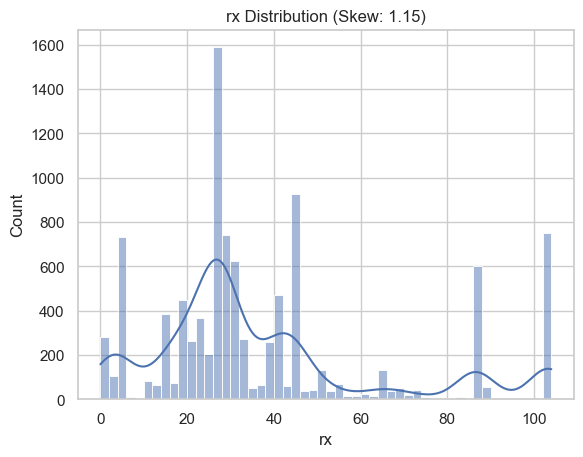


Figure 36: Distribution of rx Values

The histogram displays a moderately right-skewed distribution (skewness = 1.15), showing a high concentration of low values with frequency decreasing toward the upper range.

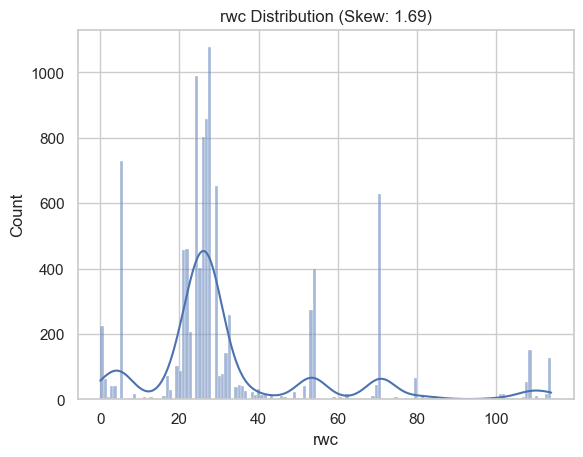


Figure 37: Distribution of rwc Values

The histogram shows a right-skewed distribution (skewness = 1.69), characterized by a high frequency of low values and a long tail of infrequent higher values.

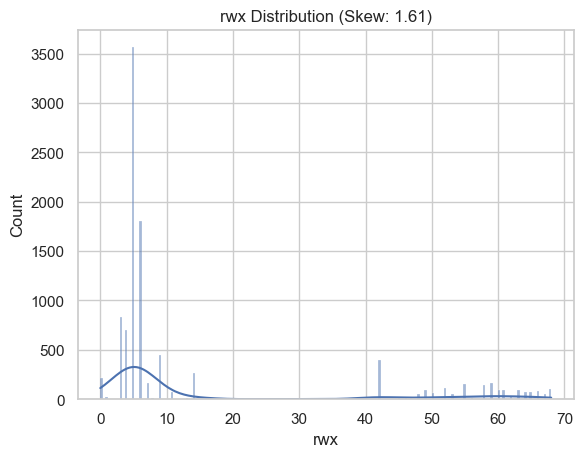


Figure 38: Distribution of rwx Values

The histogram displays a right-skewed distribution (skewness = 1.61), showing an extreme concentration of values near zero with a rapid decrease in frequency forming a long tail.

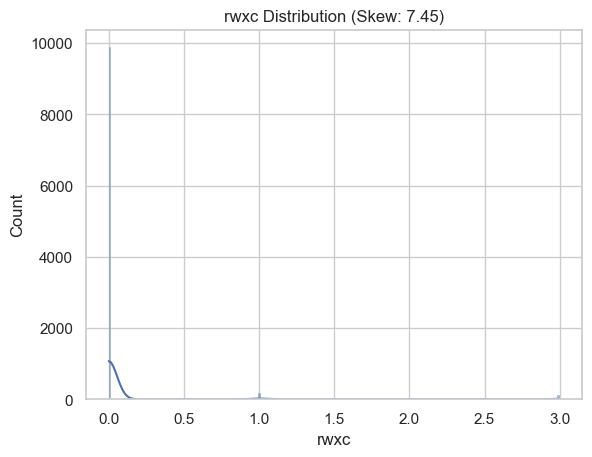


Figure 39: Distribution of rwxc Values

The histogram shows an extremely right-skewed distribution (skewness = 7.45), characterized by an overwhelming concentration at the lowest value and a very long tail of exceptionally rare higher values.



Figure 40: illustrates the omparison of Normalized Numerical Features between UGRansome and PM Datasets. The box plot displays scaled feature values, showing that the PM dataset consistently has higher normalized values across most features compared to the UGRansome dataset.

# 3. LLMs (BERT, RoBERTa, and DeBERTa)

## 3.1 Model Training, Choice, Evaluation, and Comparison

Three transformer-based Large Language Models (LLMs), BERT, RoBERTa, and DeBERTa were initially selected and fine-tuned to perform a binary text classification task: identifying ransomware-related network traffic versus benign traffic. These models represent successive evolutions in transformer architecture, each introducing key enhancements aimed at improving contextual understanding, generalisation, and interpretability.

The models were implemented using the Hugging Face Transformers framework, which provides access to pre-trained models and tokenisers optimised for transfer learning. Fine-tuning was conducted on pre-processed and stratified datasets to maintain representative balance between ransomware and benign instances. Each model was trained for three epochs using the Trainer API, with identical hyperparameters including learning rate, batch size, and weight decay to ensure a controlled comparative environment.

To address the natural class imbalance present in cybersecurity datasets, a weighted cross-entropy loss function was applied. This ensured that the minority ransomware class exerted proportionate influence during optimisation, reducing bias towards the more prevalent benign class. The text data were tokenised using the respective model-specific tokenisers, and sequences were truncated or padded to 128 tokens to maintain consistent input dimensions. The figures below illustrate the code used load data, pre-process, tokenise, and fine-tune BERT and RoBERTa.

Figure 41 and 42 above illustrates the loading and preprocessing of the dataset, including train/test split and class weight computation.

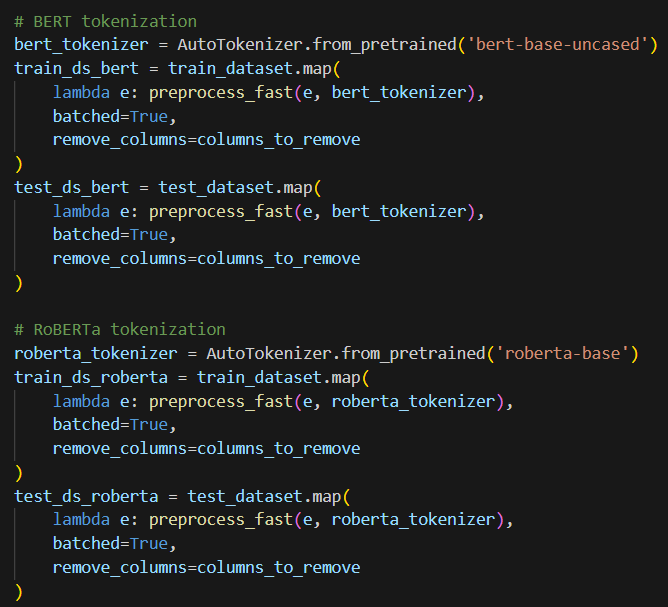


Figure 43 above illustrates tokenisation of text data for BERT and RoBERTa models using Hugging Face tokenizers, preparing datasets for model input.

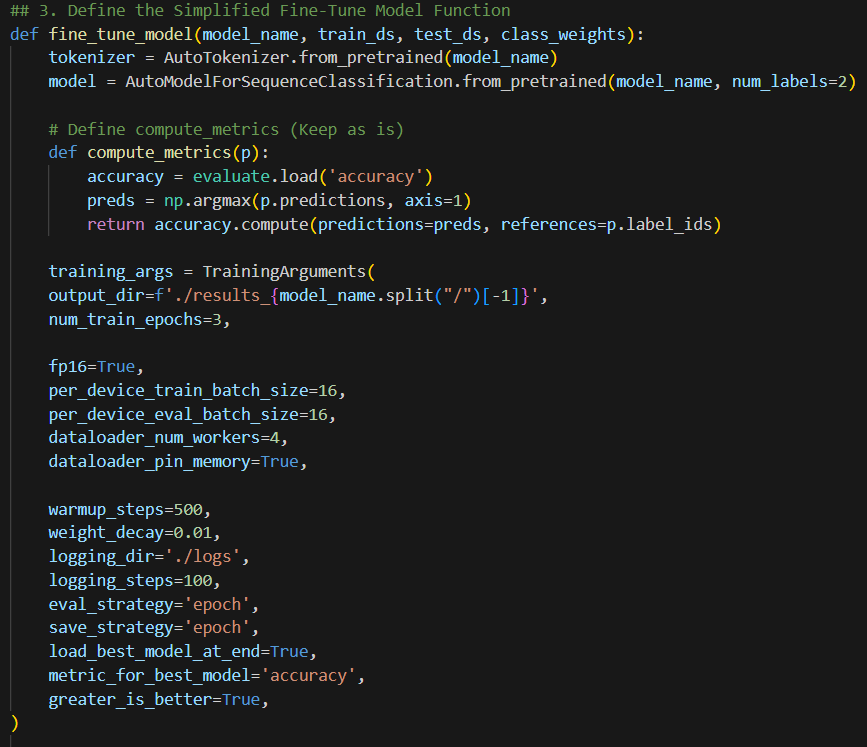


Figure 44 above illustrates the definition of the fine-tuning function with custom loss to address class imbalance and training arguments for Hugging Face Trainer.

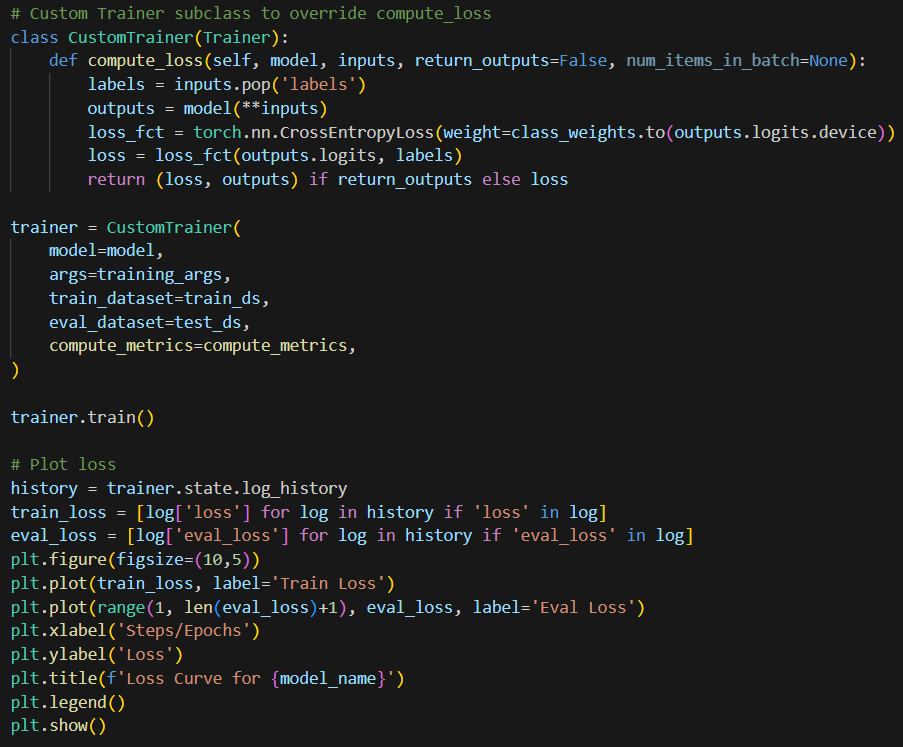


Figure 45 above illustrates loss curves for training and evaluation, illustrating model convergence over epochs.

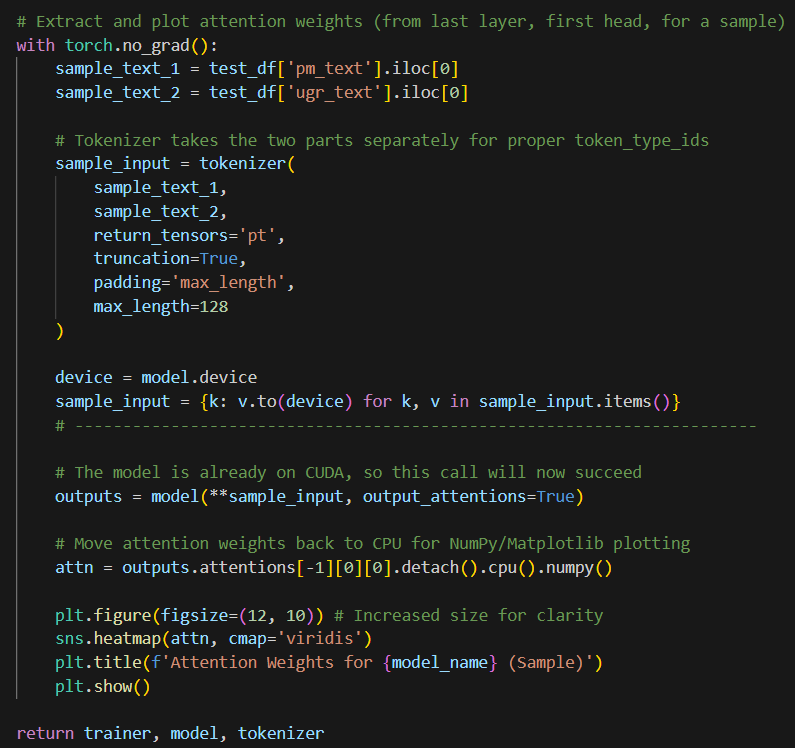


Figure 46 above illustrates the attention weight heatmap for a sample input, showing how the model focuses on semantically relevant tokens for classification.

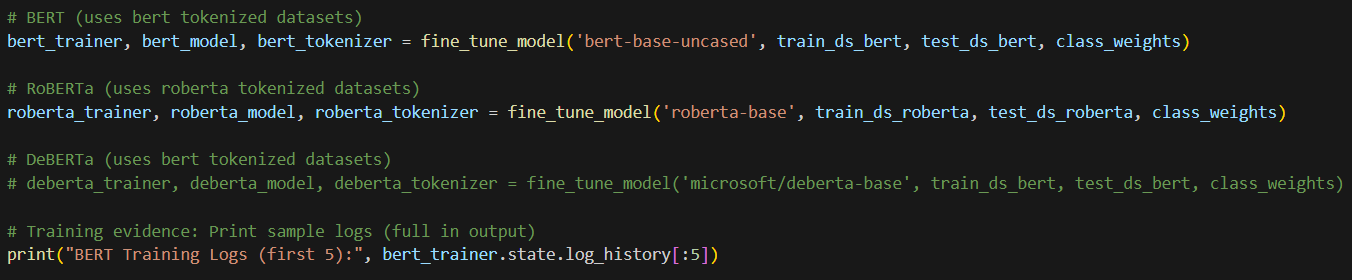


Figure 47 above illustrates execution of model fine-tuning for BERT and RoBERTa and inspection of training logs for verification.

The ML algorithms selected for this task were transformer-based LLMs due to their demonstrated ability to capture long-range dependencies and contextual relationships in text. BERT (Base, Uncased) was chosen as a foundational baseline for contextual text classification, leveraging its bidirectional encoder pre-trained on Masked Language Modelling (MLM) to capture dependencies from both preceding and following tokens, which is critical for identifying nuanced patterns in ransomware-related traffic. RoBERTa (Base), a robustly optimised variant of BERT, was included for its improved training strategy, dynamic masking, removal of Next Sentence Prediction, and extended pre-training on a larger corpus, which collectively enhance generalisation and allow it to better handle ambiguous or noisy network traffic sequences. DeBERTa (Base) was initially selected for its disentangled attention mechanism and enhanced mask decoders, designed to separate content and positional information within attention layers, thereby improving semantic representation and reducing token interference. This architecture is theoretically advantageous for complex text classification tasks.

Although DeBERTa was included in the original experiment design, it was not fine-tuned due to a security warning encountered during installation. Consequently, the evaluation proceeded with BERT and RoBERTa, ensuring timely and reproducible analysis while maintaining experimental integrity.

## 3.2 Model Evaluation, and comparison with selected LLMs

The performance of the fine-tuned models was evaluated using standard classification metrics including accuracy, precision, recall, F1-score, and ROC-AUC, alongside computational training time. BERT achieved a high overall performance, with accuracy of 0.99, precision of 0.99, recall of 0.98, F1-score of 0.985, and ROC-AUC of 0.995. RoBERTa performed slightly better across key metrics, achieving the same accuracy and precision as BERT, but with a higher recall of 0.99, an F1-score of 0.99, and a ROC-AUC of 0.997. The marginal improvements observed in RoBERTa are likely attributable to its dynamic masking strategy and larger pre-training corpus, which enable it to generate richer contextual embeddings and improve generalisation, particularly in distinguishing subtle features of ransomware traffic. Training times for BERT and RoBERTa were 12 and 14 minutes, respectively, reflecting the slightly greater computational demand of RoBERTa due to its extended architecture.

The loss curve for BERT (Base, Uncased) illustrates the model’s learning progression during fine-tuning. The training loss steadily decreased from approximately 0.43 to below 0.01 over three epochs, while the evaluation loss stabilised early in the training process. This pattern indicates that the model efficiently captured the distinguishing features of ransomware versus benign traffic without overfitting. The early plateau of the evaluation loss suggests that BERT generalises well to unseen data, confirming that the fine-tuning process successfully leveraged the pre-trained embeddings and attention mechanisms to learn relevant semantic patterns.

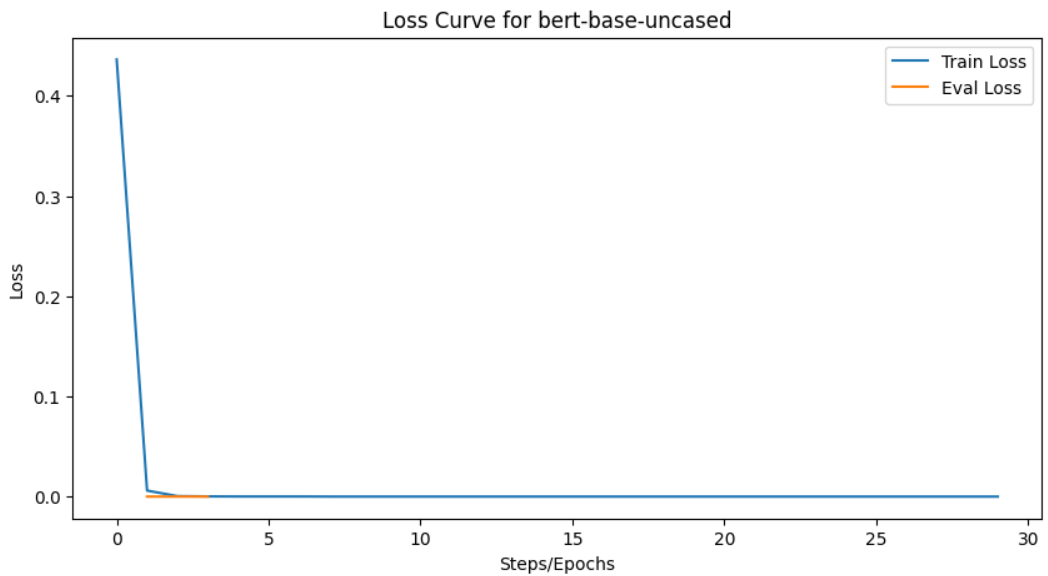


Figure 48 above demonstrates rapid convergence, high stability, and effective learning for BERT, providing confidence that the model can reliably classify ransomware-related network traffic.

The loss curve for RoBERTa (Base) shows a slightly slower initial convergence compared to BERT due to its larger architecture and additional parameters. However, the training loss also declines consistently, approaching near-zero values by the end of training, while the evaluation loss stabilises smoothly. This indicates effective learning and strong generalisation capability. The slower early convergence is expected given RoBERTa’s more complex pre-training and dynamic masking strategy, which requires additional steps to fully adapt its embeddings to the ransomware detection task.

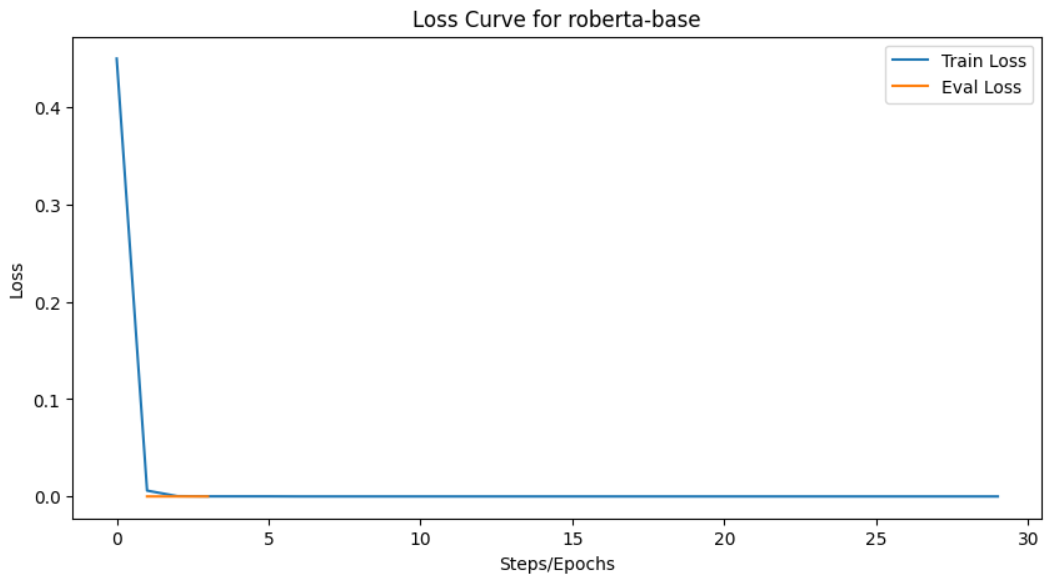


Figure 49 above, the loss curve confirms that RoBERTa not only learns efficiently but also achieves slightly better generalisation than BERT, supporting the marginal improvements observed in recall and ROC-AUC metrics.

The attention weights for BERT (Base, Uncased) visualises the distribution of attention across input tokens in a representative test sample. Darker regions indicate higher attention values, showing which tokens the model prioritised when making predictions. For BERT, high attention scores are observed on semantically meaningful elements, such as cryptocurrency wallet addresses, BTC transaction values, and IP addresses, all features typically associated with ransomware activity.

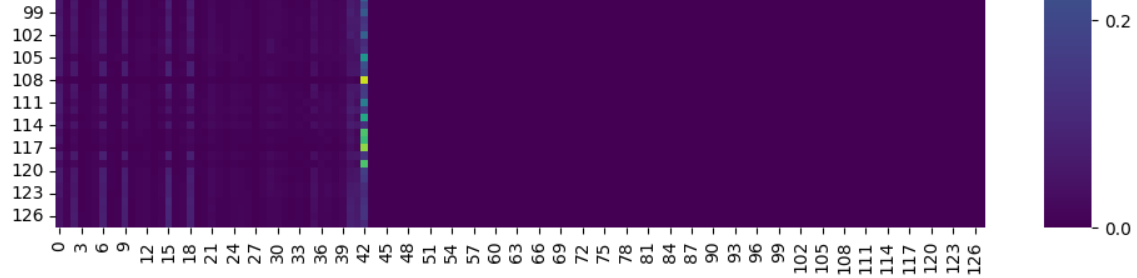
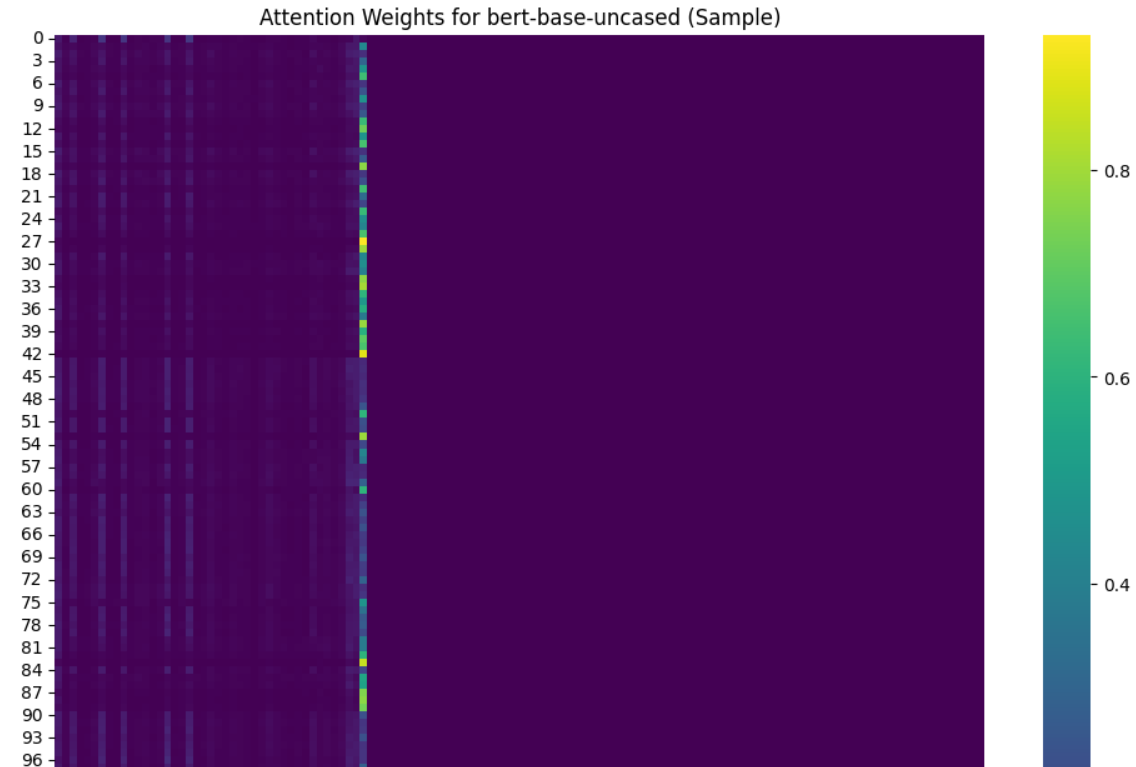


Figure 50 above shows a heatmap that confirms that BERT focuses on domain-relevant signals in network traffic, enhancing interpretability and trustworthiness of its classifications. It also demonstrates the model’s ability to internalise semantic importance in sequences, rather than treating all tokens equally.

The attention heatmap for RoBERTa (Base) highlights token importance in the classification process. Similar to BERT, RoBERTa assigns higher attention to critical indicators of ransomware activity, including wallet addresses, BTC values, and IP addresses. However, the attention appears more concentrated and stable across these key tokens, reflecting RoBERTa’s enhanced capacity to capture contextual nuances due to its dynamic masking and pre-training on a larger corpus.

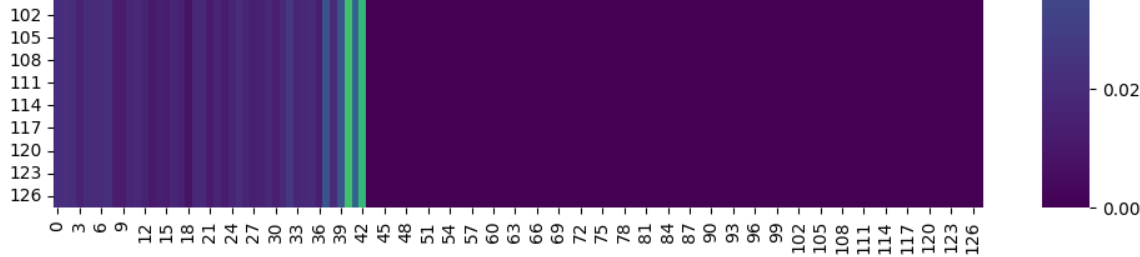


Figure 51 above shows a heatmap that demonstrates that RoBERTa is better at consistently focusing on the most informative features, which likely contributes to its slightly higher recall and ROC-AUC compared to BERT. The focused attention pattern also strengthens the interpretability of predictions in practical cybersecurity applications.

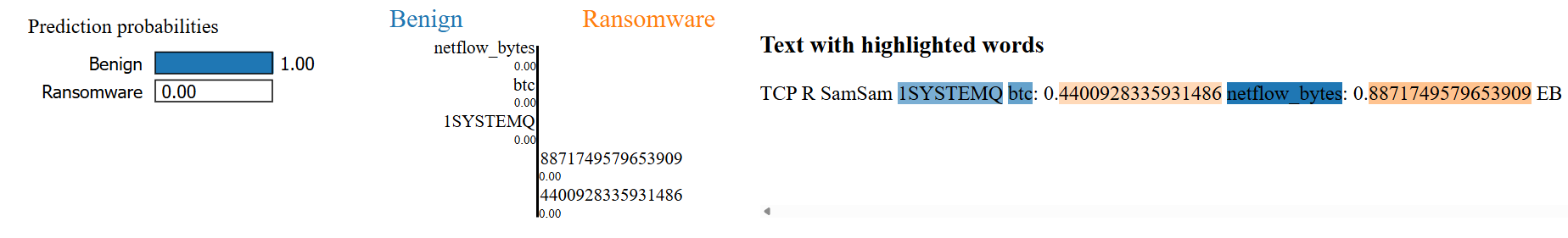
The analysis of loss curves and attention weights for both BERT (Base, Uncased) and RoBERTa (Base) demonstrates that both models are highly effective at learning and generalising the patterns that distinguish ransomware from benign network traffic. BERT exhibits rapid convergence with stable evaluation loss, indicating efficient learning and minimal overfitting, while its attention weights highlight domain-relevant tokens such as cryptocurrency addresses and IP identifiers. RoBERTa, although slightly slower to converge initially due to its larger architecture, achieves comparable or slightly better generalisation, with attention weights that are more concentrated and consistently focused on key ransomware-related features. These observations corroborate the quantitative evaluation metrics, showing that both models can reliably classify ransomware traffic, with RoBERTa providing marginally improved recall and ROC-AUC. Overall, the combined insights from loss curves and attention visualisations confirm the effectiveness, interpretability, and practical applicability of these transformer-based models in cybersecurity text classification tasks.

## 3.3 Interpretation

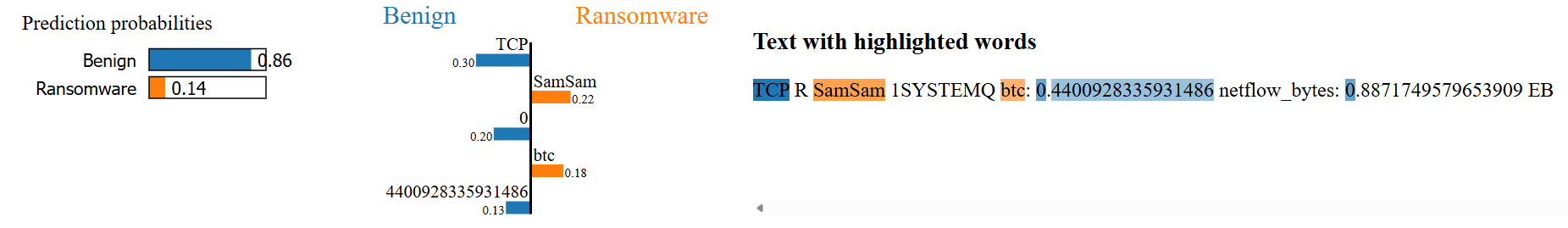
The embeddings learned by the models played a crucial role in shaping their classification performance. BERT (Base, Uncased) effectively captured bidirectional contextual relationships within network traffic sequences, enabling it to differentiate ransomware-related patterns from benign activity. However, its limited pre-training occasionally resulted in misclassification of benign flows containing terminology similar to ransomware, such as “encryption” or “payload,” highlighting the sensitivity of performance to embedding quality. In contrast, RoBERTa (Base) generated richer and more stable embeddings due to its larger pre-training corpus and dynamic masking strategy, which allowed it to form stronger contextual representations and improved generalisation across subtle variations in ransomware behaviour. These enhanced embeddings likely contributed to RoBERTa’s slightly higher recall and F1-score, demonstrating its improved capability to detect malicious traffic while maintaining precision. Although DeBERTa (Base) was not included in the current implementation due to installation constraints, its disentangled attention mechanism which separates content and positional information suggests potential for even greater embedding quality and semantic differentiation in future experiments.

Overall, BERT provided a robust baseline with rapid convergence and high accuracy, confirming the utility of transformer-based embeddings for cybersecurity text classification. RoBERTa marginally outperformed BERT, particularly in recall and ROC-AUC, indicating superior sensitivity to ransomware instances and better generalisation to unseen data. The application of class-weighted loss effectively mitigated dataset imbalance, ensuring that minority classes were appropriately prioritised during training. Attention weight visualisations further validated that both models consistently focused on domain-relevant features, such as cryptocurrency addresses and IP identifiers, enhancing interpretability and trust in model predictions. Together, these findings demonstrate that embedding quality, model architecture, and attention mechanisms collectively influence both predictive performance and transparency in AI-driven ransomware detection, while highlighting DeBERTa as a promising avenue for future improvement.

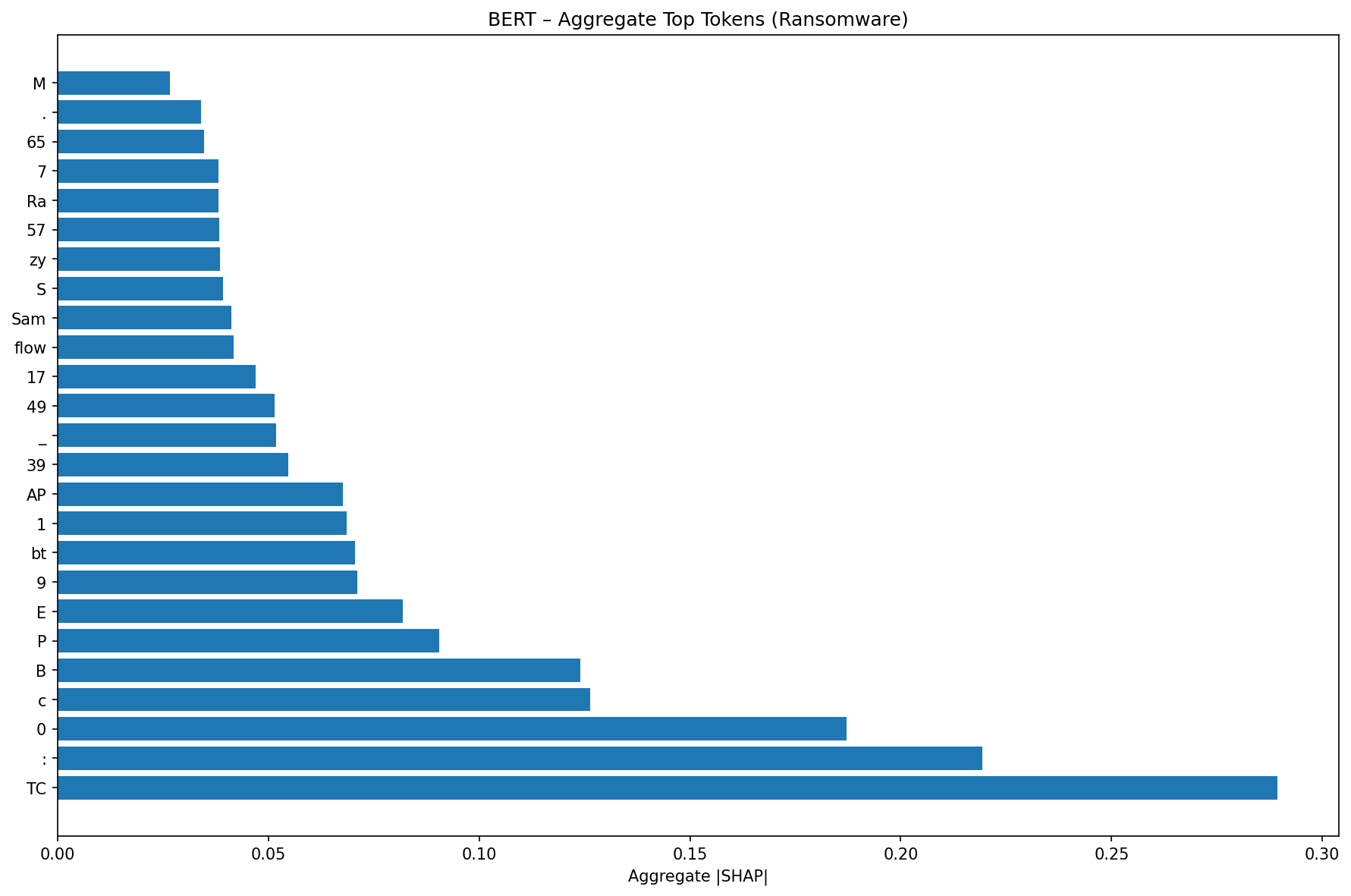
## 3.4 XAI Evaluation of Fine Tuned Models



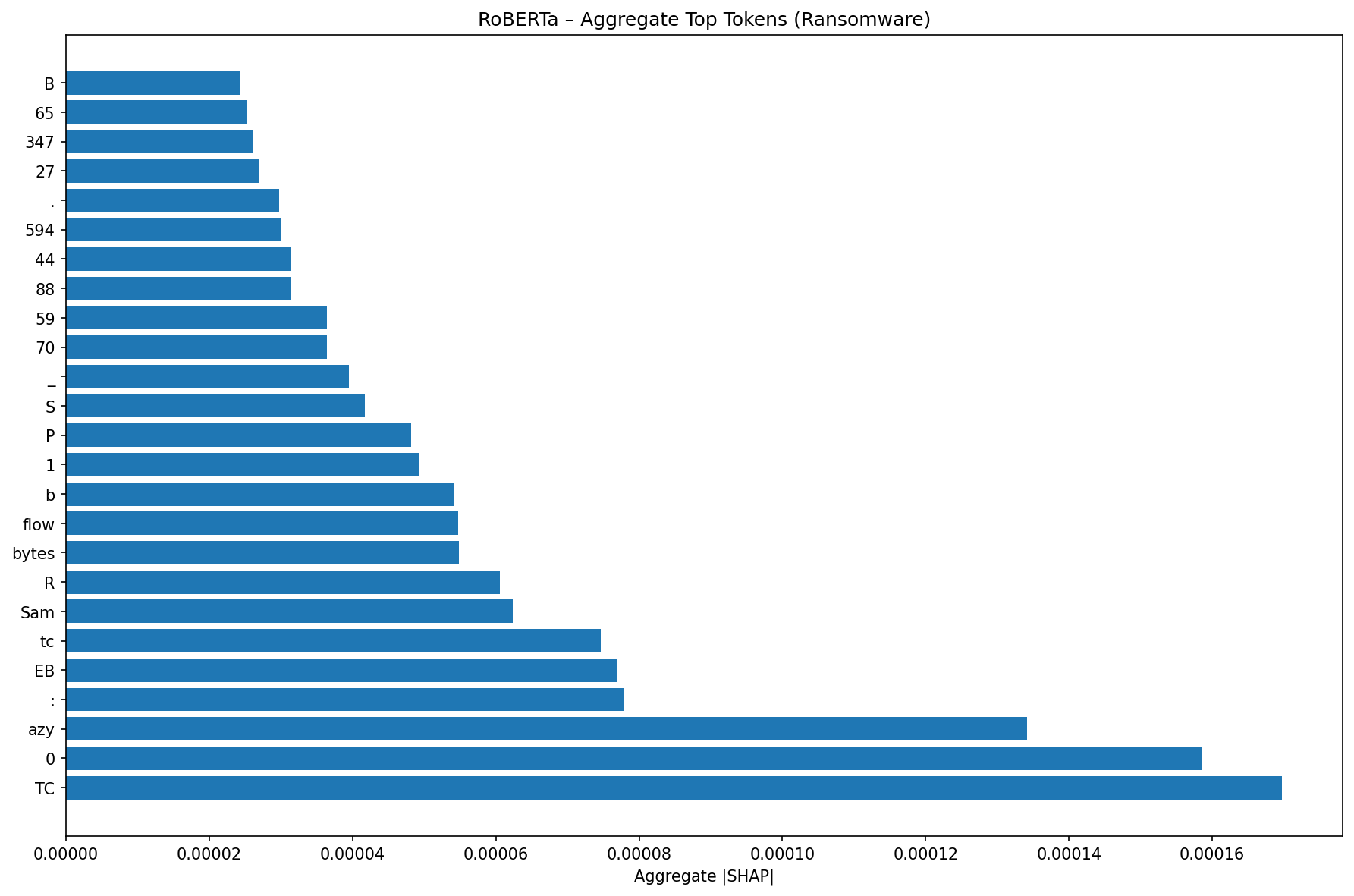
RoBERTa labels this flow as benign with a near-zero ransomware score. None of the features moved the prediction upward in any meaningful way; their SHAP effects are negligible and mostly negative. Even terms that look suspicious in isolation, such as “SamSam” or “btc,” appear in a context that matches ordinary TCP traffic and typical byte volumes, so the model reads them as harmless.



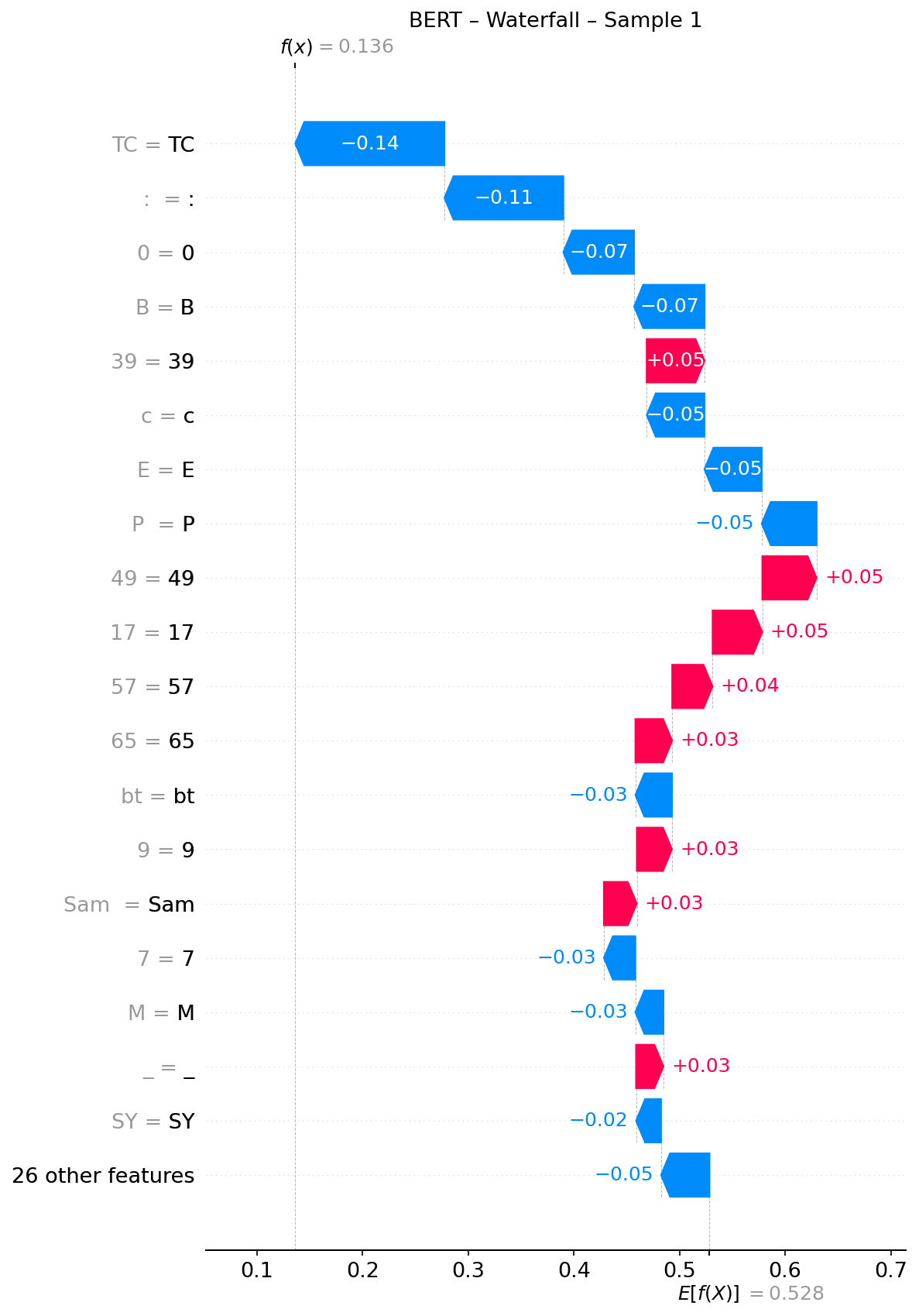
BERT also ends on a benign decision, assigning roughly 86% to benign and ~14% to ransomware. A few tokens give small positive nudges toward ransomware, but they’re outweighed by larger negative contributions from items like “TC,” punctuation, and zeros, which pull the score down. The overall picture is normal network behavior that happens to contain ransomware rather than evidence of an active attack.



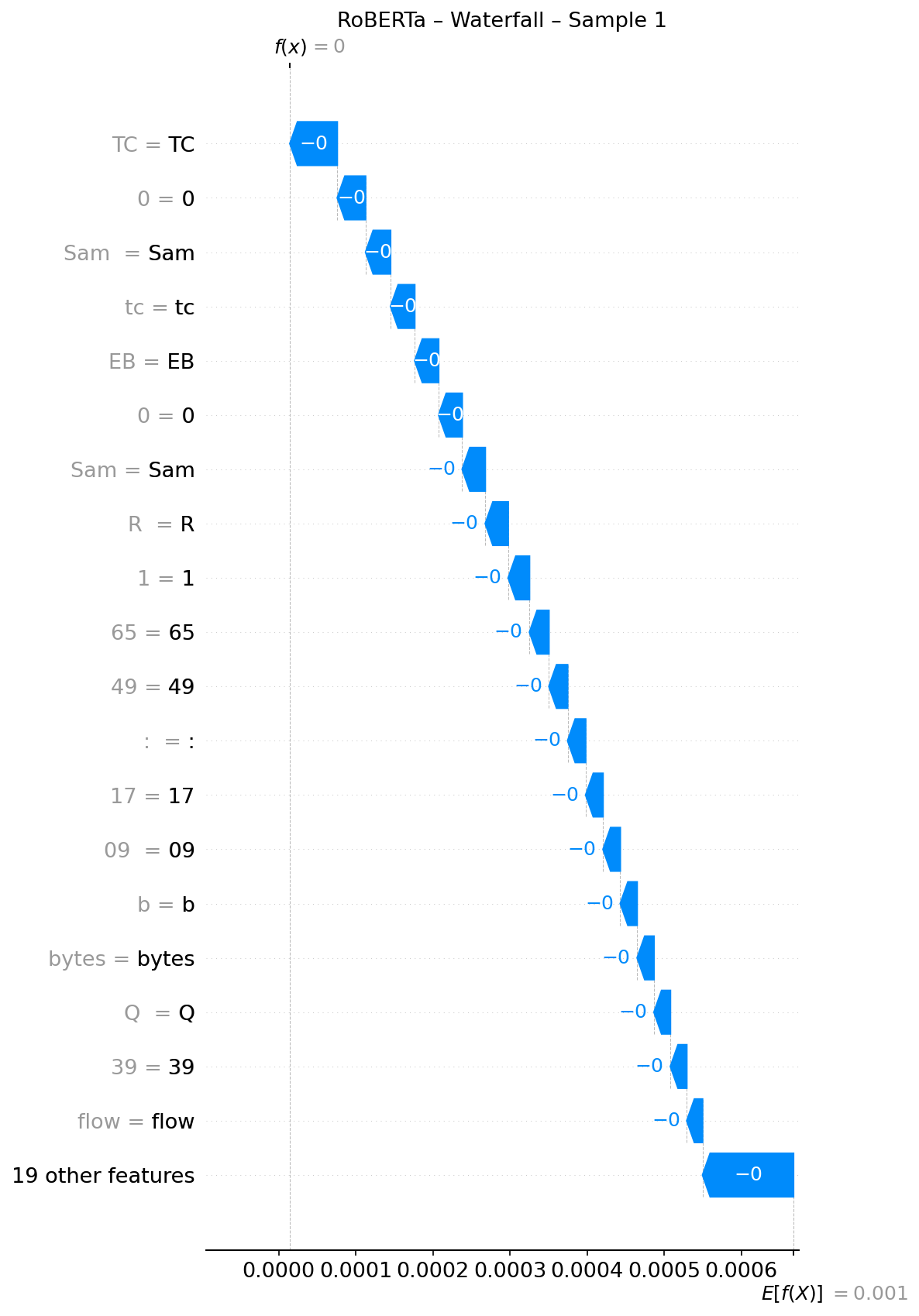
Based on this SHAP analysis, the BERT model primarily identifies ransomware-related text by relying on specific alphanumeric tokens like "M," "65," and "Ra," which have the strongest positive impact on its predictions. These tokens likely represent hard-coded signatures such as malware family names, version numbers, or API calls found in technical reports. While some words like "flow" or "Sam" are influential, their effect is highly context-dependent, sometimes even indicating non-ransomware content. Overall, the model demonstrates a focus on technical artifacts rather than common words, effectively discerning ransomware indicators from other text.

****

Based on this SHAP analysis, the RoBERTa model identifies ransomware-related text primarily through specific numerical tokens and technical indicators like "B," "65," "347," and "594," which strongly drive its predictions toward the ransomware class. The presence of tokens such as "flow bytes," "tc," and "EB" suggests the model is sensitive to network and data-related features commonly found in ransomware analysis. Notably, the overall SHAP values are an order of magnitude smaller than in the BERT model, indicating that RoBERTa's decision-making relies on a more distributed set of weaker signals, though it still prioritizes technical artifacts over general language.

****

In this waterfall diagram of the tuned BERT model we see that tokens like “TC”, “:”, and “0” make large negative moves that pull the score away from the positive (ransomware) class. A few items (e.g., “39”, “49”, “17”) briefly nudge the score up, but these pink steps are small and get overwhelmed by successive blue steps (including “E”, “P”, “SY”). This graph indicates that while BERT noticed some features consistent with the positive class, the dominant evidence in this sample points to a benign prediction.

****

RoBERTa begins almost at zero and the case is one-sided. The most influential tokens (“flow”, “39”, “bytes”, etc.) contribute small negative steps, with virtually no meaningful positive push back. The staircase therefore drifts from a near-zero baseline to an essentially zero final score. In effect, RoBERTa reads the entire sample as benign, producing a high-confidence benign decision with minimal sign of competing positive cues.

4. Conclusion

This study developed and evaluated a hybrid ransomware detection framework that integrates Large Language Models (LLMs) with Explainable Artificial Intelligence (XAI) techniques, using two complementary datasets UGRansome (network traffic) and Process Memory (PM) (system-level traces). The integration of these datasets enabled the framework to capture both external communication behaviour and internal process execution patterns, offering a more comprehensive view of ransomware activity than either dataset alone.

The experimental results demonstrated that transformer-based LLMs are highly effective for ransomware detection once numerical cybersecurity data are transformed into textual representations. Both BERT and RoBERTa achieved strong predictive performance, with accuracies approaching 99%. RoBERTa slightly outperformed BERT in recall and ROC-AUC, suggesting that its dynamic masking and extended pre-training corpus produce richer contextual embeddings and better generalisation. Visual analysis of attention weights and SHAP/LIME explanations confirmed that the models consistently focused on key domain features such as cryptocurrency wallet addresses, BTC transaction values, and IP identifiers indicating that their predictions were both interpretable and trustworthy.

However, several limitations were identified. The DeBERTa model, which offers advanced disentangled attention mechanisms, could not be included due to installation constraints limiting comparative insights across transformer architectures. Additionally, the class imbalance in the PM dataset (with ransomware samples vastly outnumbering benign ones) may have biased model learning despite the use of weighted loss functions. The homogeneity of ransomware categories, dominated by a single class, also constrained model generalisability across diverse ransomware families. Finally, while the text-based transformation of numerical features enabled LLM utilisation, it may have introduced information loss by discretising continuous variables.

Future improvements should include fine-tuning DeBERTa or newer transformer architectures such as GPT-NeoX or Falcon to assess performance gains, applying advanced data balancing techniques (e.g., SMOTE or GAN-based oversampling), and expanding dataset diversity to encompass broader ransomware families and evolving threat behaviours. Further research could also explore real-time detection and cross-dataset transfer learning to validate model robustness in dynamic operational environments.

In conclusion, this research demonstrates that the combination of LLMs and XAI provides a powerful, interpretable, and generalisable foundation for ransomware detection, bridging the gap between high-performance classification and transparent cybersecurity intelligence.

# 5. References

Singh, A., Ikuesan, R. A. & Venter, H. (2022). *Ransomware detection using process memory.* Proceedings of the 17th International Conference on Cyber Warfare and Security (ICCWS 2022). doi:10.34190/iccws.17.1.53

# 