A Comparative Analysis of Machine Learning Algorithms for Website Traffic Classification from Network Packets

Matthew Berthoud, Lake Bradford, Justin Cresent, Will Katabian

William & Mary

Williamsburg, Virginia, USA

ABSTRACT

Accurately Identifying web traffic destination and origins is crucial for the efficiency of a network. This project explores the potential of machine learning in reference to web traffic classification based on the analysis of network packets. We monitored and analyzed web traffic data from ChatGPT, Blackboard, and Linkedin, with the objective of building models which will be able to predict the web traffic origin of a specific packet. The collection of data was performed using wireshark, then the data was reformatted to eliminate bias and get more accurate results. Using the collected data we then trained four models which had varying levels of accuracy, Logistic Regression (56%), K-Nearest Neighbors (77%), Random Forest (78%), and finally a neural network (80%). This project shows the importance of machine learning within the field of network traffic analysis as automaton is much more efficient and precise compared to manually examining web traffic, especially in a scale as large as the internet. One example of our project's significance is that this can be crucial data analysis for network administrators and security professionals, whom would examine the network for malicious traffic.

KEYWORDS

Network Traffic, Machine Learning, Web Traffic Classification, Network Packets, Data Analysis, Neural Networks, Random Forest, K-Nearest Neighbors, Logistic Regression

1 INTRODUCTION

The ability to monitor and analyze network traffic is crucial for the efficiency and security of a network [4]. It helps to manage the overall network performance, detect and prevent malicious activities, and ensure the network is operating as intended. The present day internet is composed of a vast variety of diverse web traffic, of which requires a more sophisticated approach to analyze and classify network traffic[2]. This project investigates a variety of machine learning algorithms to see which most accurately and effectively classify web traffic based on data from a set of captured network packets.

Traffic classification is very significant in practice, if done accurately and effectively [5]. For reasons mentioned previously, the observed capabilities of this and other projects can be crucial for network administrators and security professionals, whom would examine the network for malicious traffic.

For this project, many sets of packet data were collected from three popular websites: ChatGPT, Blackboard, and Linkedin. These sets were then merged into a single dataset, which was then split into training and testing sets. Four models were then trained and tested on this data: Logistic Regression, K-Nearest Neighbors, Random Forest, and Neural Network [3].

A thorough evaluation of the models was conducted through testing and validation sets, hyperparameter optimization employing cross-validation and grid search, and computing accuracy and Macro F1 scores. The results showed that the Neural Network model acheived the highest accuracy of 80% and Macro F1 score of BLANK. These scores display the model's ability to accurately classify the selected websites based on captured network packet data.

Overall, the significance of this project is seen in its use of Machine Learning and the extensions of these applications into the discipline of network traffic analysis. This practice has potential to be utilized in real-world applications, and many sources proving it already is [2].

2 PROPOSED METHOD

This section presents the methodology for how this project goes about capturing network packet data, analyzing the data, and classifying it based on its website of origin.

2.1 Capturing Data

2.1.1 Wireshark.

2.1.2 Data Collection. To commence the collection of our data, we open Wireshark and select a network connection. In this project we established a connection the College of William and Mary's network interface (en0). In order to isolate the web traffic in order to specifically collect packets that we are analyzing, the capture filter functionality is utilized. The website of interest is then visited in the web browser

which causes Wireshark to begin collecting the packets being exchanged. After collecting sufficient packets, the data is then converted into a Comma Seperated Values (CSV). This process of data collection is performed for Blackboard, LinkedIn, and ChatGPT resulting in the creation of three distinct files.

2.2 Data Preparation and Processing

To maintain normalized representation in all the collected datasets, a process of data harmonization was applied after collecting the necessary data. The dataset with the minimum number of observations is identified and accounted for when truncating the remaining two datasets. Subsequently, the three datasets are combined and rearranged to minimize bias while training each of our models. This process ensures a balanced yet unbaised dataset that maximizes the effectivness of our models.

2.2.1 Feature Selection.

2.2.2 Data Formatting.

2.3 Employed models

2.3.1 Logistic Regression (Baseline).

2.3.2 K-Nearest Neighbors.

2.3.3 Random Forest.

2.3.4 Neural Network.

3 EVALUATION

This section comprehensively evaluates our models and their performance on the dataset and validating sets. Later, goes on to determine the model which best classified our selected websites based on network packet data.

3.1 Dataset

3.2 Evaluation Metrics

Accuracy and Macro F1 score were the primary metrics used to evaluate each model. These metrics are applicable to all models and allow for a direct comparison for each of their performances. Two other metrics: precision and recall, help to provide a more detailed understanding of the model's performance on a class-to-class basis.

The equation for **Accuracy**:

$$Accuracy = \frac{TP + TN}{TS}$$

The equation for **Macro F1**:

$$MacroF1 = \frac{1}{n} \sum_{1}^{n} \frac{2 * Precision_i * Recall_i}{Precision_i + Recall_i}$$

The equation for **Recall**:

$$Recall = \frac{TP}{TP + FN}$$

The equation for **Precision**:

$$Accuracy = \frac{TP}{TP + FP}$$

Accuracy is the score that rates the effectivness of a model at classification overall (ie. how many it predicited correctly out of the total set).

Macro F1 is the harmonic average of the of precision and recall for every class, and then taking the average of these calculated scores accross all classes.

Recall is the calculation of how many the model correctly classified within a class (ie. how many truly belong in that class in the entire dataset). Precision is the calculation of how many the model correctly classified within a class (ie. can be thought of as class accuracy).

3.3 Model Comparison and Evaluation Results

4 DISCUSSION & FUTURE WORK

4.1 Analysis of Results

4.2 Future Work

5 CONCLUSION

REFERENCES

[1] Wireshark Foundation. 2024. Wireshark.. https://www.wireshark.org/.
[2] Kay Boldt, Kenneth B. Kent, and Rainer Herpers. 2020. Investigation of encrypted and obfuscated network traffic utilizing machine learning. In *Proceedings of the 30th Annual International Conference on Computer Science and Software Engineering* (Toronto, Ontario, Canada) *(CASCON '20).* IBM Corp., USA, 43–52.
[3] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
[4] Hiroshi Tsunoda and Glenn Mansfield Keeni. 2012. Security by simple network traffic monitoring. In *Proceedings of the Fifth International Conference on Security of Information and Networks* (Jaipur, India) *(SIN '12).* Association for Computing Machinery, New York, NY, USA, 201–204. https://doi.org/10.1145/2388576.2388608
[5] Jun Zhang, Xiao Chen, Yang Xiang, Wanlei Zhou, and Jie Wu. 2015. Robust network traffic classification. *IEEE/ACM Trans. Netw.* 23, 4 (aug 2015), 1257–1270. https://doi.org/10.1109/TNET.2014.2320577

[4] [2] [5] [3] [1]