Erkennung bösartiger URLs mit großen Sprachmodellen

Ivaylo Krumov
University of Luxembourg
Email: ivaylo.krumov.001@student.uni.lu

Dieser Bericht wurde unter der Aufsicht von:
Salima Lamsiyah
University of Luxembourg
Email: salima.lamsiyah@uni.lu

1. Einleitung

In der heutigen vernetzten Welt findet ein schneller Informationsaustausch über das Internet statt. Mit dem technologischen Fortschritt ändern sich jedoch auch die Methoden der Cyberkriminellen. Eine der ältesten und häufigsten Cyberattacken ist die Verwendung bösartiger URLs, die darauf abzielen, Benutzer zu täuschen und unbefugten Zugang zu vertraulichen Informationen zu erhalten. Dieses Projekt untersucht die Anwendung großer Sprachmodelle (LLMs) zur Erkennung bösartiger URLs und konzentriert sich dabei auf neuronale Netzwerkarchitekturen, die große Mengen von Textdaten verwenden, um Antworten auf der Grundlage gegebener Eingaben zu generieren. Diese Fähigkeit könnte genutzt werden, um Muster in Webadressen zu erkennen, die auf böswillige Absichten hinweisen.

2. Wissenschaftliches Ergebnis

2.1. Anforderungen

Der wissenschaftliche Beitrag zielt auf die Beantwortung der Frage ab: Wie können wir große Sprachmodelle verwenden, um bösartige URLs zu erkennen?“ Es enthält Erläuterungen zu relevanten Konzepten wie der Feinabstimmung von vortrainierten LLMs, der Datenvorverarbeitung und der Bewertung der Modellleistung. Das Dokument stützt sich auf einschlägige akademische und wissenschaftliche Quellen, um einen Überblick über den Bereich der Cybersicherheit und der Aufklärung von Cyberbedrohungen zu geben, und erörtert maschinelles Lernen, die Verarbeitung natürlicher Sprache und spezifische Merkmale relevanter Modelle.

2.2. Gestaltung

Der wissenschaftliche Beitrag ist ein Text, der den technischen Beitrag ergänzt, indem er den Kontext zu dessen Ideen und Umsetzung liefert. Er erforscht relevante Themen, fasst Informationen zusammen und formuliert eine schlüssige Antwort auf die wissenschaftliche Frage. Es wird auf mehrere wissenschaftliche Quellen verwiesen, um sicherzustellen, dass die präsentierten Informationen korrekt und fundiert sind.

2.3. Produktion

Die Untersuchung beginnt mit neuronalen Netzen und konzentriert sich auf transformatorbasierte Architekturen wie BERT und GPT. Diese Modelle werden mit großen Mengen an Textdaten trainiert und können für bestimmte Aufgaben feinabgestimmt werden. Für die Erkennung bösartiger URLs wird das Modell auf Datensätzen trainiert, die sowohl gutartige als auch bösartige URLs enthalten, und lernt dabei strukturelle Merkmale, die auf bösartige Absichten hinweisen. Die Feinabstimmung umfasst die Vorverarbeitung der Daten, die Merkmalsextraktion und die Abstimmung der Hyperparameter zur Optimierung der Modellleistung.

Fallstudien zeigen die Wirksamkeit verschiedener maschineller Lerntechniken bei der Erkennung bösartiger URLs. Saleem Raja et al. zeigten beispielsweise eine hohe Genauigkeit bei der Verwendung lexikalischer Merkmale und maschineller Lernklassifizierer wie Random Forest. Eine andere Studie von Abdi und Wenjuan untersuchte Faltungsneuronale Netze zur Erkennung bösartiger URLs und erzielte eine hohe Genauigkeit. Diese Beispiele in Verbindung mit theoretischem Wissen zeigen, dass LLMs bösartige Muster in URLs effektiv analysieren und identifizieren können.

2.4. Bewertung

Der wissenschaftliche Beitrag liefert detaillierte Erklärungen der relevanten Konzepte und kontextualisiert das theoretische Wissen für den Anwendungsfall der Erkennung bösartiger URLs. Er befasst sich erfolgreich mit der wissenschaftlichen Frage und liefert eine konkrete Antwort, die den Leser darauf vorbereitet, den technischen Beitrag zu erkunden.

3. Technisches Ergebnis

3.1. Anforderungen

Die technische Leistung umfasst die Implementierung von mindestens zwei großen Sprachmodellen, die auf die Klassifizierung von URLs zugeschnitten sind. Die Modelle sollten anhand eines öffentlich zugänglichen Datensatzes feinabgestimmt und anhand von Metriken wie Genauigkeit, F1-Score und Präzision bewertet werden. Der Code sollte gut dokumentiert und in Jupyter-Notebooks organisiert sein, um die Lesbarkeit und Benutzerfreundlichkeit zu gewährleisten.

3.2. Gestaltung

Der technische Beitrag ist in vier Jupyter-Notebooks gegliedert, die jeweils die Datenvorverarbeitung und das Modelltraining für die URL-Klassifizierung behandeln. Die gewählten Modelle sind ein neuronales Feed-Forward-Netzwerk, BERT, DistilBERT und GPT-2. Das Design gewährleistet Flexibilität beim Experimentieren mit verschiedenen Modellen, Parametern und Techniken.

3.3. Produktion

Der verwendete Datensatz besteht aus 651.190 URLs, die als gutartig, defacement, phishing oder malware gekennzeichnet sind. Jedes Modell wurde an einer Stichprobe von 10.000 URLs trainiert und bewertet, wobei die Daten vorverarbeitet wurden, u. a. durch Konvertierung der Bezeichnungen und Tokenisierung.

Das neuronale Feed-Forward-Netzwerk erreichte eine durchschnittliche Genauigkeit von 99.12%. Das BERT-Modell zeigte eine starke Leistung mit hoher Genauigkeit und geringem Validierungsverlust. DistilBERT, eine destillierte Form von BERT, erreichte eine ähnliche Leistung, jedoch mit kürzeren Trainingszeiten. Der Versuch, GPT-2 mit Hilfe der Transformer Reinforcement Learning Bibliothek zu implementieren, war aufgrund technischer Herausforderungen nicht erfolgreich.

3.4. Bewertung

Der technische Beitrag zeigt erfolgreich die Durchführbarkeit und Wirksamkeit der Verwendung von LLMs für die URL-Klassifizierung. Trotz der unvollständigen Implementierung des Verstärkungslernansatzes erfüllt das Ergebnis die festgelegten Anforderungen und liefert die erwarteten Ergebnisse.

4. Schlussfolgerung

Die in diesem Bericht vorgestellten Forschungen und Analysen haben das Potenzial und die Praktikabilität der Verwendung großer Sprachmodelle zur Erkennung bösartiger URLs aufgezeigt. Die wissenschaftliche Untersuchung deckte wesentliche Konzepte ab, und der technische Beitrag implementierte und bewertete mehrere LLMs, die auf die Klassifizierung von URLs zugeschnitten sind. Das Projekt zeigt erfolgreich, dass die Integration von theoretischem Wissen mit praktischen Implementierungen, mit angemessener Feinabstimmung und Optimierung, LLMs zu leistungsfähigen Werkzeugen in der Cybersicherheit machen kann, um die Auswirkungen von sich entwickelnden Bedrohungen zu mindern.