RESUMEN

El phishing es un tipo de ataque cibernético cuyo objetivo es engañar a los usuarios, generalmente a través de páginas web aparentemente benignas. Actualmente, una de las formas más comunes de detectar estas páginas de phishing es mediante el análisis de su contenido. Esto implica analizar el texto de las páginas web y posteriormente examinar ese contenido con algoritmos de Deep Learning (DL). Según el estado del arte, el texto se introduce de forma secuencial en los algoritmos de DL, es decir, sin considerar el orden o el significado de las palabras. Este método, por lo tanto, ignora la riqueza semántica inherente a las relaciones entre las palabras. La innovación de este estudio propone un modelo que emplea el Procesamiento de Lenguaje Natural (NLP) y algoritmos Transformer de DL para detectar ataques de phishing basándose en el texto extraído de páginas web sospechosas. En este trabajo, se utiliza la Metodología de Descubrimiento de Conocimiento en Bases de Datos (KDD) para realizar un análisis comparativo de cuatro modelos basados en la arquitectura Transformer, junto con NLP, para identificar ataques de phishing a partir del texto contenido en dichos ataques. Se inicia con la selección y preprocesamiento de un conjunto de datos obtenido del sitio PhishTank, que incluye textos de páginas de phishing y textos de páginas legítimas (ham). Posteriormente, se aplican técnicas de limpieza, tokenización y lematización para preparar los datos para el análisis. Para maximizar las características semánticas y sintácticas del contenido de la página web, se emplea la capa de incrustación de Keras con GloVe. Finalmente, estos datos se procesan con cuatro modelos Transformer para determinar qué algoritmo funciona mejor: BERT, GPT, T5 y XLNet. Las evaluaciones experimentales demostraron que el modelo BERT, con un 93.34%, obtuvo la mejor precisión. Como complemento a nuestra investigación, se desarrolló una extensión para el navegador Chrome que permite a los usuarios verificar si una página es de phishing o no.

Palabras clave: Phishing, Aprendizaje Profundo, Procesamiento de Lenguaje Natural, Transformer, BERT, GPT, T5, XLNet, GloVe.

ABSTRACT

Phishing is a form of cyber attack whose purpose is to deceive users, generally through websites that appear benign. Currently, one of the most common ways to detect these phishing pages is based on their content. That is, analyzing the text on web pages and then analyzing that content with deep learning (DL) algorithms. According to the state of the art, this text is entered consecutively into the DL algorithms, regardless of the order or meaning of the words. In this way, this method loses the richness of semantics between the relationship between words. The innovation of our study is to propose a model that uses Natural Language Processing (NLP) and DL Transformer algorithms to detect phishing attacks based on the text extracted from suspicious web pages. In this work, the Knowledge Discovery Methodology in Databases (KDD) is used to perform a comparative analysis of four models based on the Transformer architecture together with NLP to detect phishing attacks on the text contained in these attacks. We begin with selecting and pre-processing a data set obtained from the Phishload site, including text from phishing pages and text from legitimate (ham) pages. Then, we use cleansing techniques, tokenization, and lemmatization to prepare them for analysis. To maximize the semantic and syntactic features of the web page content, we utilize the Keras embedding layer with GloVe. Finally, this data was processed with four Transformer models to determine which algorithm works best: BERT, GPT, T5, and XLnet. Experimental evaluations showed that the 93.03% BERT model obtained the best accuracy. An extension was made to complement our research in the Chrome browser, allowing users to determine if a page is phishing.

Keywords: Phishing, Deep Learning, Natural Language Processing, Transformer, BERT, GPT, T5, XLnet, GloVe.