Ministry of Higher Education and
Scientific Research
University of Mosul
College of Computer Science and
Mathematics
Department of Computer Science

A Model to Detect Arabic Phishing E-mails Using NLP

**A Thesis Submitted to the Council of the College of
Computer Science and Mathematics
University of Mosul
as a Partial Fulfillment of Requirements
for the Degree of Master of Science
in
Computer Science**

By

Rian Sharafaldeen Yahya Al-Yozbaki

Supervised by

Prof. Dr. Mafaz Mohsin Khalil Alanezi

2023 A.D. **1445 A.H.**

Abstract

Phishing is a risk associated with social engineering that exploits the gullibility of uninformed Internet users to trick them into revealing personal information. Attackers and phishers assume the guise of real Internet users. Where phishers try to access the victim's accounts without their permission to steal sensitive data, the victim's identity, and other personal information. Phishing emails are a form of phishing where a phisher sends an email to the recipient using a fake email address to trick them into reading the email. This enables the phisher to influence the user and take advantage of their personal data.

Many researchers worked on finding solutions and methods to detect phishing emails, most of which were related to the English language, and because the Arabic language is distinguished from other languages in terms of the formulation of sentences and the multiplicity of meanings of words. There is a dearth of research related to the Arabic language, especially those that depend on natural language processing techniques. Also, the researchers faced a major problem, through the lack of a dataset of phishing emails in the Arabic language.

The proposed model Rian AntiPhishing (RAPH) was created using Python language and its libraries. The model relies on natural language processing techniques to analyze the Arabic content of the emails and perform multiple comparisons at the word, root and phrase levels most commonly used in the phishing emails. The model works by directly connecting to any Gmail account or any account affiliated with the University of Mosul email domain, and performs analysis, comparison, and processing in real time. Also, a dataset of Arabic phishing emails and legitimate emails was created that includes 1250 emails, as well as creating two datasets for comparison: The first included a list of the most common phishing words, and the second included a list of the most commonly used phrases in phishing.

To ensure the validation of the effectiveness of the configured model and to know the impact of phishing, an integrated phishing system was created (for research purposes), a number of phishing campaigns were conducted on the same Gmail account, and the generated model was run for phishing detection. The verified results of the model showed its effectiveness in accurately diagnosing phishing messages, after a series of experiments for

individual comparisons. At the end, it turned out that the best results can be achieved when implementing the system using all comparisons together. The best results percentages were phishing detection (98.4%) with legitimate detection (92.5 %) and phishing detection (96.0%) with legitimate detection (99.5 %) at use multiple values for comparison variables.

وزارة التعليم العالي والبحث العلمي
جامعة الموصل
كلية علوم الحاسوب والرياضيات
قسم علوم الحاسوب

نموذج لإكتشاف رسائل البريد الإلكتروني الاحتيالية باللغة العربية باستخدام معالجة اللغات الطبيعية

رسالة مقدمة
الى مجلس كلية علوم الحاسوب والرياضيات في جامعة الموصل
كجزء من متطلبات نيل شهادة ماجستير علوم في
علوم الحاسوب
من قبل

ريان شرف الدين يحيى اليوزبكي

بإشراف
الأستاذ الدكتور مفاز محسن خليل العنزي

1445 هـ
2023 م

الملخص

التصيد الاحتيالي هو أحد المخاطر المرتبطة بالهندسة الاجتماعية التي تستغل سذاجة مستخدمي الإنترنت غير المطلعين لخداعهم في الكشف عن معلومات شخصية. يتبنى المهاجمون والمخادعون ستار مستخدمي الإنترنت الحقيقيين. حيث يحاول المحتالون الوصول إلى حسابات الضحية دون إذن منه لسرقة البيانات الحساسة وهوية الضحية وغيرها من المعلومات الشخصية. رسائل البريد الإلكتروني الخادعة هي أحد أشكال التصيد الاحتيالي حيث يرسل المخادع بريدًا إلكترونيًا إلى المستلم باستخدام عنوان بريد إلكتروني مزيف لخداعه لقراءة البريد الإلكتروني. يمكّن ذلك المخادع من التأثير على المستخدم والاستفادة من بياناته الشخصية.

عمل العديد من الباحثين على إيجاد حلول وطرق للكشف عن رسائل البريد الإلكتروني التصيدية ، والتي كان معظمها متعلقًا باللغة الإنجليزية، ولأن اللغة العربية تتميز عن غيرها من اللغات من حيث صياغة الجمل وتعدد معاني الكلمات ، فهناك ندرة في الأبحاث المتعلقة باللغة العربية وخاصة تلك التي تعتمد على تقنيات معالجة اللغة الطبيعية. ايضا واجهت الباحثين مشكلة كبيرة وذلك من خلال عدم توفر مجموعة بيانات لرسائل البريد الالكتروني للتصيد الاحتيالي باللغة العربية.

تم إنشاء النموذج بإستخدام لغة بايثون والمكتبات التابعة لها. حيث يعتمد النموذج على تقنيات معالجة اللغة الطبيعية لتحليل المحتوى العربي لرسائل البريد الإلكتروني وإجراء عمليات مقارنة متعددة على مستوى الكلمة والجذر والجمل الأكثر استخدامًا في رسائل البريد الإلكتروني المخادعة. يعمل النموذج من خلال الاتصال المباشر بأي حساب جيميل أو أي حساب تابع لمخدم ايميلات جامعة الموصل, ويقوم بإجراء التحليل والمقارنة والمعالجة في الوقت الفعلي.أيضًا تم إنشاء مجموعة بيانات لرسائل التصيد الإلكتروني العربية ورسائل البريد الإلكتروني الشرعية تتضمن 1250 بريدًا إلكترونيًا، وكذلك انشاء مجموعتي بيانات للمقارنة تضمنت الأولى قائمة بالكلمات الاكثر شيوعا في التصيد الاحتيالي، وتضمنت الثانية قائمة بأكثر الجمل استخداما في التصيد الاحتيالي.

لضمان التحقق من فعالية النموذج الذي تم تكوينه ومعرفة تأثير التصيد الاحتيالي ، تم إنشاء نظام تصيد متكامل (لأغراض البحث) ، وتم إجراء عدد من حملات التصيد على نفس حساب الجيميل ، وتشغيل النموذج الذي تم إنشاؤه لاكتشاف التصيد.

أظهرت النتائج التي تم التحقق منها للنموذج فعاليته في التشخيص الدقيق لرسائل التصيد ، بعد سلسلة من التجارب للمقارنات الفردية والجماعية. حيث اتضح أنه يمكن تحقيق أفضل النتائج عند تنفيذ النظام باستخدام جميع المقارنات معًا. كانت أفضل النسب المئوية هي لاكتشاف رسائل

التصيد بنسبة (98.4٪) مقابل اكتشاف رسائل شرعية بنسبة (92.5٪), ونسبة اكتشاف رسائل التصيد (96.0٪) مع اكتشاف رسائل شرعية بنسبة (99.5٪) عند استخدام قيم متعددة لمتغيرات المقارنة.