Phishing via Instant Messaging Systems on Social Media Sites Using Artificial Intelligence Techniques in the Field of Social Engineering

Dr. Majd Ahmad Ali*

(Received 17 / 12 / 2023. Accepted 17 / 3 / 2024)

□ ABSTRACT □

Phishing is a highly effective form of cybercrime that enables criminals to deceive users and steal sensitive data. With the growing number of internet users, particularly on social media platforms, there has been a substantial increase in publicly available personal information, making these users more susceptible to phishing attacks. This research provides a detailed explanation of the use of artificial intelligence techniques in social engineering, especially the application of deep learning and transformers in textual content analysis and context understanding. Based on this analysis, an attacker can identify the victim's area of interest to form the basis of their attack and steal user data by sending a customized message tailored for each victim based on their interests using instant messaging systems on the same platforms. The results demonstrated that phishing attacks using messages customized to the victim's area of interest have a success rate of up to 70%.

Keywords: Internet crimes - phishing - deep learning - natural language processing - transformers.

* **Assistant Professor, Department of Artificial Intelligence, Faculty of Informatics Engineering, Tishreen university, Syria. majdahmadali@gmail.com**

التصيد الاحتيالي عبر أنظمة التراسل الفوري في مواقع التواصل الاجتماعي باستخدام تقنيات الذكاء الصنعي في مجال الهندسة الاجتماعية

د. مجد احمد علي*

(تاريخ الإيداع 17 / 12 / 2023. قُبِل للنشر في 17/ 3 / 2024)

□ ملخّص □

يعد التصيد الاحتيالي شكلاً فعالاً للغاية من الجرائم الإلكترونية التي تمكّن المجرمين من خداع المستخدمين وسرقة البيانات المهمة. ومع النمو المتزايد لمستخدمي الإنترنت ولا سيما مواقع التواصل الاجتماعي فقد تزايد كثيرا عدد مشاركات المعلومات الشخصية، ما يجعل مستخدمي هذه المواقع أكثر عرضةً لهجمات التصيد الاحتيالي. يقدم هذا البحث شرحاً تفصيليا لاستخدام تقنيات الذكاء الاصطناعي في الهندسة الاجتماعية، ولا سيما استخدام التعلم العميق والمتحولات لتحليل المنشورات النصية وفهمها بشكل سياقي، والتي على أساسها يمكن لمنفذ الهجوم تحديد مجال اهتمام الضحية ليشكل قاعدة إطلاق هجمته باتجاهها بهدف سرقة بياناتها المهمة من خلال إرسال رسالة مخصصة حسب مجال اهتمامها باستخدام أنظمة التراسل الفوري في ذات المواقع. بينت النتائج أن التصيد الاحتيالي باستخدام الرسائل المخصصة حسب مجال اهتمام الضحية تصل نسبة نجاحه إلى 70%.

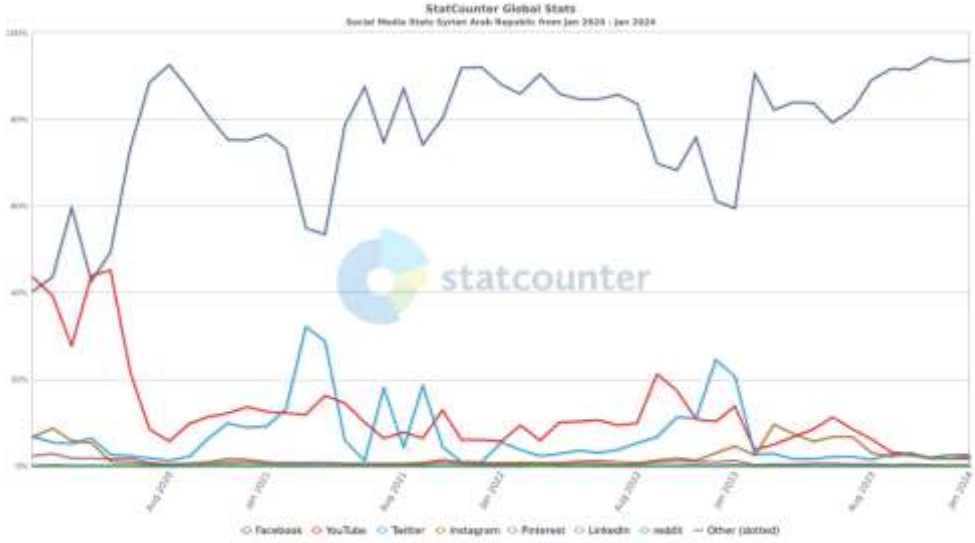
الكلمات المفتاحية: جرائم الإنترنت –التصيد الاحتيالي – تعلّم عميق – معالجة لغات طبيعية – المتحولات.

* **مدرس – قسم الذكاء الصنعي – كلية الهندسة المعلوماتية – جامعة تشرين-اللاذقية- سورية. majdahmadali@gmail.com**

مقدمة:

يقضي كثير من مستخدمي الإنترنت مدداً زمنية طويلة في تصفّح مواقع التواصل الاجتماعي على اختلاف أنواعها، للحصول على المعلومات وللترفيه ولتوثيق لحظاتهم المهمة، أو للتعبير عن آرائهم تجاه سلعة أو موضوع معين، فضلاً عن أنّ هذه المواقع أصبحت أساسية للكثير من الشركات وأصحاب الأعمال في التسويق لخدماتهم، وتطوير أعمالهم تبعاً لتوجهات الزبائن.

نلاحظ من خلال مراقبة احصائيات استخدام السوريين لمواقع الإنترنت بشكل عام و مواقع التواصل الاجتماعي بشكل خاص الموضحة في الشكل(1) والتي يقدمها موقع StatCounter في الفترة ما بين كانون ثاني 2020 و كانون ثاني 2024 [1]، أن غالبية مستخدمي الإنترنت في سورية هم من رواد مواقع التواصل الاجتماعي و لا سيما موقع فيسبوك و منصة إكس (تويتر سابقاً)، و هذا ما يجعل غالبية مستخدمي الإنترنت في سوريا عرضة لهجمات التصيد الاحتيالي Phishing القائم على الهندسة الاجتماعية Social Engineering.



الشكل 1 – إحصائيات موقع2 statcounter لمستخدمي منصات التواصل الاجتماعي في سورية في الفترة بين كانون ثاني 2020 وكانون ثاني 2024 [1]

مع تزايد شعبية منصات التواصل الاجتماعي أصبح استخدامها جزءاً أساسياً للتواصل اليومي لمعظم الأفراد والشركات، لذلك أصبح من الأهمية بمكان أن يتم استخدامها بعناية ومسؤولية واتخاذ كافة الإجراءات اللازمة لحماية البيانات الخاصة بحسابات مستخدمي هذه المنصات ولا سيما الوعي بالتصيد الاحتيالي.

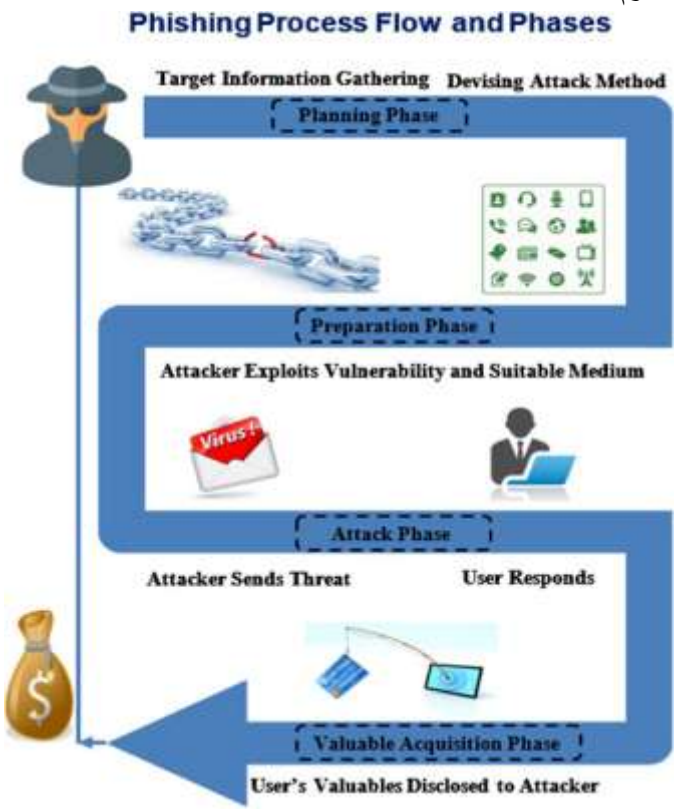
يعتبر التصيد الاحتيالي من أكثر التهديدات المرتبطة بأمن وسائل التواصل الاجتماعي Social Media Cybersecurity ويمكن تعريفه على أنه محاولة خادعة للحصول على معلومات حساسة، مثل كلمات المرور أو التفاصيل الشخصية أو تفاصيل بطاقة الائتمان وطرق الدفع وغيرها، من خلال انتحال هوية كيانات حقيقية أو إنشاء ملفات تعريف مزيفة [2].

يعتمد مهاجمو التصيد الاحتيالي على الهندسة الاجتماعية لخداع الأشخاص ودفعهم للتخلي عن بياناتهم الحساسة. والطريقة الأكثر شيوعًا للقيام بذلك هي من خلال رسائل البريد الإلكتروني والرسائل الفورية [3].

وفقًا لأحدث تقرير صادر عن مجموعة عمل مكافحة التصيد الاحتيالي Anti-Phishing Working Group (APWG)، فقد زاد عدد هجمات التصيد الاحتيالي بأكثر من 150% سنويًا منذ بداية عام 2019. ولذلك، تم تطوير أساليب جديدة للكشف عن الرسائل التصيدية وكشف نقاط الضعف فيها [4].

من ناحية أخرى يمكن أن نعرّف التصيد الاحتيالي بأنه هجوم اجتماعي تقني، حيث يستهدف المهاجم أشياء ثمينة محددة من خلال استغلال ثغرة أمنية موجودة لتمرير تهديد محدد عبر وسيط محدد إلى نظام الضحية، باستخدام حيل الهندسة الاجتماعية أو بعض التقنيات الأخرى لإقناع الضحية باتخاذ إجراء محدد يسبب أنواعًا مختلفة من الأضرار [5].

يوضح الشكل (2) مخطط سير أي عملية تصيد احتيالي وهو يحتوي على أربع مراحل، حيث أنه، في معظم الهجمات، تبدأ عملية التصيد الاحتيالي بمرحلة التخطيط وتصميم الهجوم ويتم ذلك من خلال جمع معلومات حول الهدف واختيار أسلوب الهجوم الملائم بناء على تلك المعلومات. المرحلة الثانية هي مرحلة الإعداد، حيث يقوم المهاجم بالبحث عن نقاط الضعف التي يمكن استغلالها لخداع الضحية. يقوم المهاجم بتنفيذ هجومه في المرحلة الثالثة وينتظر الرد من الضحية. بمجرد استجابة الهدف تبدأ المرحلة الرابعة حيث يحاول المهاجم الحصول على أكبر قدر ممكن من المعلومات والبيانات الخاصة بالضحية، وكخطوة أخيرة يقوم المهاجم بمسح آثار الهجوم (حظر حساب الضحية على سبيل المثال) في عملية الهجوم [5].

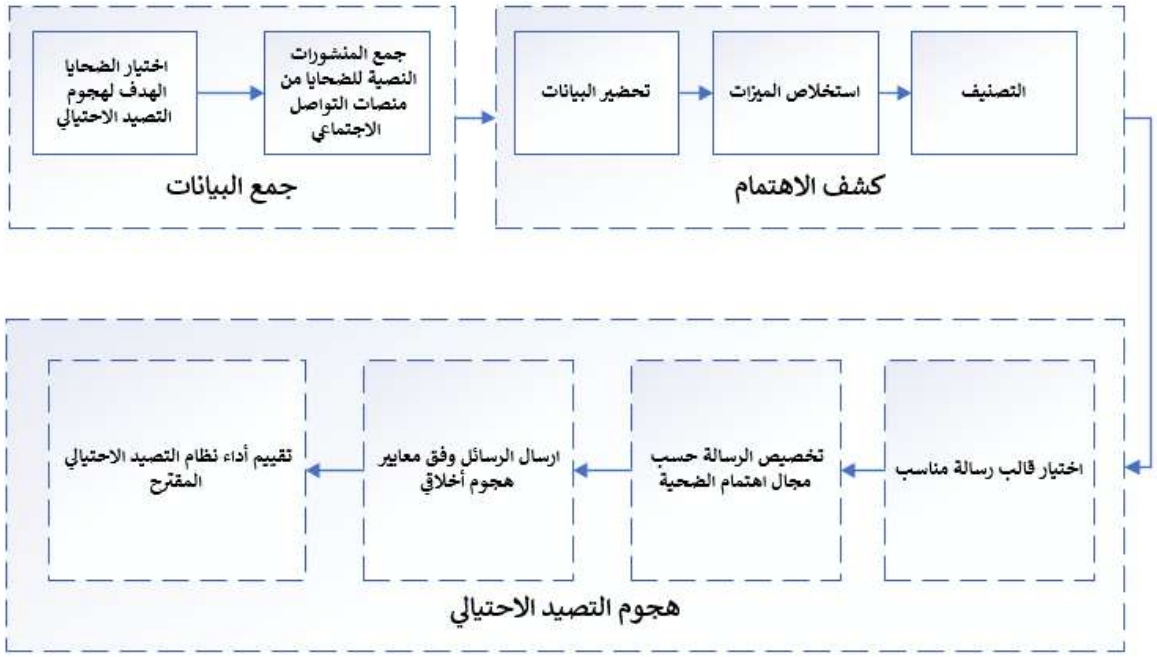


الشكل (2): المخطط العام لسير عملية التصيد الاحتيالي [5]

لتوضيح عملية التصيد الاحتيالي المذكورة سابقاً نعرض المثال التالي، قد يرسل المهاجم رسالة احتيالية إلى مستخدم إنترنت يتظاهر بأنه من بنك الضحية، ويطلب من المستخدم تأكيد تفاصيل الحساب المصرفي، وإلا فقد يتم تعليق الحساب. قد يعتقد المستخدم أن هذه الرسالة حقيقية لأن المهاجم يستخدم نفس العناصر الرسومية والعلامات التجارية

والألوان الخاصة ببنكه. فيرسل المعلومات المطلوبة طوعاً إلى المهاجم الذي سيستخدمها لأغراض ضارة مختلفة مثل سحب الأموال أو الابتزاز أو ارتكاب المزيد من عمليات الاحتيال [5].

يقترح بحثنا نظامًا لتخصيص هجمات التصيد الاحتيالي بناءً على مجال اهتمام الضحايا المستهدفين باستخدام تقنيات معالجة اللغات الطبيعية Natural Language Processing (NLP)، ودراسة فاعلية هذا النظام الموضح في الشكل (3) بالاعتماد على ردود أفعال الضحايا في تجربة هجوم أخلاقي Ethical Attack Experiment.



الشكل (3): مخطط النظام المقترح

إحدى الطرق الشائعة للتعامل مع مشكلة اكتشاف الاهتمامات هي نمذجة الموضوع Topic Modeling، وهو أسلوب تعليم غير خاضع للإشراف Unsupervised Learning لاكتشاف موضوع مستند ما، لكن هذه الآلية لا تعمل بشكل جيد مع النصوص القصيرة بسبب صعوبة استخلاص السياق منها. كما أنه من الصعب توصيف المواضيع التي اكتشفها هذا النهج. ونتيجة لذلك، قررنا استخدام تعلم الآلة الخاضع للإشراف Supervised Learning وتدريب مصنف على أداء المهمة المذكورة.

في هذه المقالة، سنقوم بمقارنة أداء مصنفات تعلم الآلة التقليدية Machin Learning وخوارزميات التعلم العميق Deep Learning والمتحولات Transformers في تصنيف موضوع المنشورات النصية على وسائل التواصل الاجتماعي كوسيلة لاكتشاف مجال اهتمام المستخدم.

أهمية البحث وأهدافه:

تزداد هجمات التصيد الاحتيالي ويزداد تعقيدها بشكل كبير ما يجعل منها تهديداً أمنياً جدياً. يهدف البحث إلى تحسين الأمن السيبراني Cybersecurity للأفراد والمنظمات من خلال تقديم آلية دقيقة لفهم إحدى أدوات التصيد الاحتيالي الحديثة القائمة على استخدام الذكاء الصنعي وكشف نقاط الضعف في أنظمة الحماية الحالية بهدف تلافيها واتخاذ تدابير مضادة أكثر فعالية.

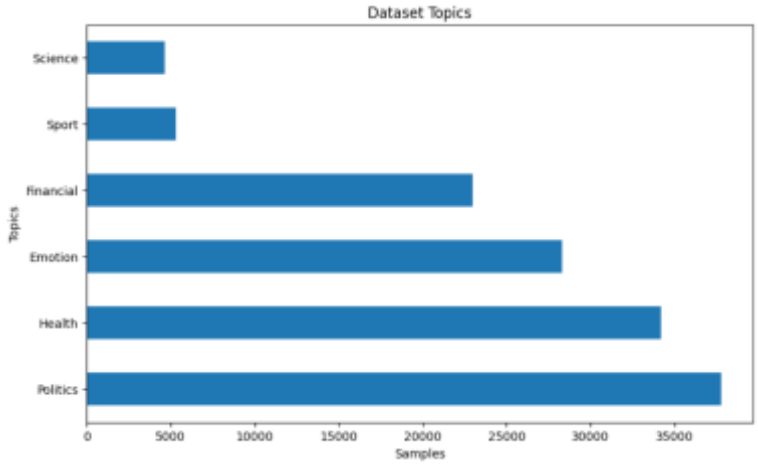
تعتمد تقنيات كشف و منع التصيد الاحتيالي التقليدية على النظم الخبيرة القائمة على القواعد Rule Based Expert System وبعض أنظمة تعلم الآلة التي تبقى محدودة في قدرتها على كشف أساليب الهجوم المتطورة، حيث تسمح تقنيات الهندسة الاجتماعية بصياغة رسائل شخصية مقنعة وتساعد الطبيعة الديناميكية للهجمات على تجاوز أنظمة الحماية من هجمات التصيد الاحتيالي المستخدمة حالياً.

تأتي أهمية البحث من قدرة خوارزميات تعلم الآلة المتقدمة وتقنيات معالجة اللغات الطبيعية على تحليل محتوى المنشورات النصية في وسائل التواصل الاجتماعي وتحديد مجال اهتمام المستخدم مما يمكن المهاجم من إنشاء رسائل تصيد مخصصة لكل مستخدم بما يتناسب مع مجال اهتمامه. والذي يعتبر بدوره عاملاً حاسماً في زيادة احتمالية نجاح الهجوم.

طرائق البحث ومواده:

مجموعة البيانات المستخدمة Dataset:

بداية لا بد من شرح مجموعة بيانات التدريب المستخدمة حيث استخدمنا مجموعة بيانات مفتوحة المصدر Open Source تسمى "Topic Classification dataset ". مجموعة البيانات هذه متاحة على منصة Kaggle [6]، وهي تقدم حوالي 136000 منشوراً نصياً باللغة الإنكليزية تم جمعها من مصادر مختلفة، مثل منشورات وتعليقات وسائل التواصل الاجتماعي والأخبار والمقالات. تنتمي كل عينة إلى واحد من الأصناف التالية: السياسة والصحة ومجتمع والاقتصاد والرياضة والعلوم [6].



الشكل (4): مخطط التوزع التكراري للأصناف في مجموعة البيانات[6]

تصنيف النصوص:

يُقصد بتصنيف النصوص تجميع النصوص أو تصنيفها إلى فئات حسب مضمون النص، وله العديد من التطبيقات منها استرجاع المعلومات Information Retrieval وعنونة الموضوع Topic Labeling وتحليل المشاعر Sentiment Analysis وتلخيص النصوص Text Summarization والإجابة عن الأسئلة Question Answering (QA) والهندسة الاجتماعية.

تنوَّعت الخوارزميات والطرق المستخدمة في تصنيف النصوص بدءاً من طرق التصنيف المعتمدة على القواعد، تليها خوارزميات تعلم الآلة التقليدية، ثم الخوارزميات المعتمدة على شبكات التعلم العميق بما فيها المتحولات [7].

تتم عملية تصنيف النصوص وفق عدة مراحل بداية بإجراء معالجة مسبقة للنصوص Text Pre-processing، ثم استخلاص السمات (الميزات) Feature Extraction وتشكيل أشعة السمات لنصوص الدخل، يليها تدريب المصنف، وأخيراً اختبار دقة التصنيف وتقييم النتائج. يبين الشكل (5) مراحل عملية التصنيف.

الشكل (5): مراحل عملية تصنيف النصوص

المعالجة المسبقة للنصوص:

قمنا بدايةً بإزالة العينات التي تحتوي على أكثر من 300 كلمة نظرًا لأن هدفنا هو تصنيف منشورات منصات التواصل الاجتماعي التي عادة ما تكون منشورات قصيرة حيث تم اختيار 300 كلمة كطول أعظمي للعينة بناءً على التحليل الاحصائي لأطوال العينات في مجموعة البيانات المستخدمة في البحث. بعد ذلك استخدمنا خطوات المعالجة المسبقة التالية:

- تقييس النص Text Normalization: تحويل النص إلى شكل قياسي قد يختلف عن شكله الأصلي.
- إزالة علامات الترقيم Punctuation Removal: حذف علامات الترقيم من النص، مثل الفواصل والنقاط وعلامات الاستفهام.
- إزالة كلمات التوقف Stop Words Removal: إزالة الكلمات التي تظهر بشكل متكرر وليس لها أهمية كبيرة في النص. على سبيل المثال الكلمات مثل (and, of, the, a, an) هي كلمات توقف في اللغة الإنجليزية.
- إزالة واسمات HTML HTML Tags Removal: حذف واسمات HTML من النص، وهي عبارة عن رموز تشير إلى بنية وتنظيم اللغة المكتوبة في صفحة الويب.
- إزالة الأحرف الخاصة Special Characters Removal: حذف الأحرف الخاصة من النص. مثل (,* ,# $,&) والرموز التعبيرية Emojis، التي لا تشكل جزءًا من الأبجدية القياسية.
- إصلاح الاختصارات Contractions Fix: الاختصارات هي كلمات أو مجموعات من الكلمات التي يتم اختصارها عن طريق إسقاط الحروف واستبدالها بفاصلة عليا. على سبيل المثال، (I'll) هي اختصار لـ (I will). يشير إصلاح الاختصار إلى عملية إعادة التعابير إلى شكلها الأصلي.

تمثيل الميزات:

تمثيل الميزات أو استخراجها في مجال معالجة اللغة الطبيعية هو عملية تحويل البيانات النصية إلى تمثيلات عددية، وهو أحد الأجزاء الأساسية لأي تطبيق معالجة لغات طبيعية NLP application نظرًا لأن الحاسبات لا تستطيع فهم الأحرف والكلمات والجمل.

استخدمنا طريقتين رئيسيتين لاستخراج الميزات من البيانات النصية:

- التوزين وفق تردد المصطلح وتردد المستند العكسي Term Frequency –Inverse Document Frequency(TF–IDF):

وهو أحد طرق استخلاص الميزات ضمن نموذج حقيبة الكلمات Bag of Word (BoW) الأكثر شيوعاً، حيث يتم فيها تمثيل الجملة أو الوثيقة بشعاع ذو عدد سمات مساوٍ لمجموع كل الكلمات الفريدة الموجودة في مجموعة بيانات

التدريب، ويوجد عدة طرق لتوزين الكلمات ضمن هذا النهج وأهمها TF-IDF حيث يعكس TF-IDF مدى أهمية الكلمة في النص [8].

$$TF - IDF = TF * IDF \quad (1)$$

TF: هو نسبة عدد الكلمات الموجودة في الجملة إلى طول الجملة:

$$TF = \frac{\text{the count of a word present in a sentence}}{the\ total\ number\ of\ words\ in\ the\ sentece} \quad (2)$$

IDF للكلمة: هو لوغاريتم LOG نسبة إجمالي عدد الصفوف في مستند معين إلى عدد الصفوف التي تحوي هذه الكلمة:

$$IDF = \log\left(\frac{N}{n}\right) \quad (3)$$

حيث N هو إجمالي عدد الصفوف في المستند و n هو عدد الصفوف التي تحوي الكلمة.

- تضمين الكلمات Word Embedding:

يعالج هذا النهج مشكلة عدم أخذ سياق النص بالحسبان في نموذج حقيبة الكلمات، حيث يتم في هذا النهج تمثيل كل كلمة بشعاع عددي وتأخذ الكلمات المتشابهة في المعنى تمثيلات متقاربة من بعضها في فضاء الأشعة، ويكون شعاع الجملة مجموع أشعة الكلمات المكونة لها. أشهر النماذج المستخدمة في هذا النهج: word2vec, fasttext.

كلا النموذجين word2vec وfasttext يستخدمان شبكة عصبونية بسيطة Shallow Neural Network إذ يتم تدريب النموذج بهدف تعديل أوزانه و استخدام هذه الأوزان في تشكيل تمثيل شعاعي للكلمات.

أحد أفضل نماذج تضمين الكلمات هو نموذج word2vec، وهو نموذج للتعلم العميق يستخدم جميع كلمات مجموعة النص لإنشاء متجه لكل كلمة بناءً على سياقها [8].

1- خوارزميات تصنيف النصوص:

قمنا باستخدام المصنفات التالية:

1. مصنفات تعلم الآلة التقليدية: هناك مجموعة متنوعة من خوارزميات تعلم الآلة التي يمكن استخدامها لتصنيف النص، وقد وقع الاختيار على المصنفات التالية:
   - آلة متجهات الدعم Support Vector Machine (SVM).
   - الانحدار التدريجي العشوائي Stochastic Gradient Descent (SGD).
   - الانحدار اللوجستي Logistic Regression (LR).
   - المصنف البايزي Naive Bayes (NB).
2. خوارزميات التعلم العميق: تم استخدام نوع من الشبكات العصبونية التكرارية Recurrent Neural Networks يدعى شبكات الذاكرة الطويلة قصيرة الأمد Long Short-Term Memory (LSTM)، حيث تتمتع هذه الشبكات ببنية مؤلفة من بوابات لاتخاذ القرار حول المعلومات التي سيتم الاحتفاظ بها أو إهمالها، وتشتهر بأدائها الجيّد عند التعامل مع النصوص.
3. المتحولات: تعرف المتحولات بأنها نماذج تعلم عميق ذات بنية مؤلفة من مرمز Encoder ومفكك ترميز Decoder. تعتمد المتحولات على آلية الانتباه الذاتي Self-Attention Mechanism لحساب تمثيلات الدخل والخرج بدلاً من استخدام الوصلات التسلسلية الموجودة في الشبكات العصبونية التكرارية.

2- قالب رسالة التصيد الاحتيالي:

بناء على ما توصلنا إليه في جزأ النظام الخاص بتحديد مجال اهتمام الضحية المزعومة من خلال مهمة تصنيف النصوص الواردة في الصفحة الشخصية يمكن تصميم رسالة تصيد مقنعة.

لقد اخترنا القالب التالي لاستخدامه في تجربة التصيد الاحتيالي:

Are you interested in {Topic}? Do you want to stay updated on the latest news, trends, and tips on {Topic}? Do you want to win a {Prize}?

We're offering you a chance follow our {Topic} page and enter a draw to win a {Prize}

Our {Topic} page delivers the latest and most relevant information on {Topic} to you. You will get access to exclusive articles, videos, podcasts, and more.

To follow our page and win your {Prize}, all you have to do is click on the link below:

{Phishing Link}

الشكل (6): قالب رسالة التصيد الاحتيالي المستخدم

يمكننا تخصيص هذا القالب بسهولة عن طريق استبدال {name} و{topic} و{prize} لكل هدف. ويتم تتبع الرابط المضمن باستخدام أداة تسمى Grabify. يمكن لهذه الأداة تتبع عنوان بروتوكول الإنترنت IP الخاص بالهدف، والمتصفح، ونظام التشغيل، واسم المضيف، ومزود خدمة الإنترنت [9].

3- سيناريو اختبار نظام التصيد الاحتيالي المقترح:

أبلغنا المجموعة المستهدفة بالغرض من الدراسة وأكدنا لهم أنه لن يتم جمع أي معلومات حساسة أو إلحاق أي ضرر من خلال سيناريو هجوم التصيد الاحتيالي المصمم وأن الهدف من الهجوم هو هدف بحثي. بعد ذلك، طلبنا معرف حساباتهم الشخصية على منصة X.

بعد الحصول على معرف الحسابات الشخصية على منصة X لأهداف الهجوم، استخدمنا واجهة برمجة التطبيقات Application Programming Interface (API) التي تقدمها المنصة لجمع منشوراتهم اعتمادًا على لغة البرمجة Python ومكتبة تسمى Tweepy. توفر مكتبة Tweepy واجهة سهلة الاستخدام للوصول إلى واجهة برمجة التطبيقات (API) لأداء مهام مختلفة، مثل البحث عن المنشورات والوصول إلى معلومات المستخدم [10].

استخدمنا النموذج الذي أعطى النتائج الأفضل لتصنيف منشورات الأهداف واستخراج اهتماماتهم. بعد ذلك، قمنا بتخصيص الرسائل لكل مستخدم حسب مجال اهتمامه الرئيسي. أخيراً قمنا بمراسلة كل هدف وإرسال النسخة المخصصة له من الرسالة وانتظرنا الرد.

النتائج والمناقشة:

1- نتائج نظام تحديد اهتمام مستخدمي منصات التواصل الاجتماعي:

يوجد العديد من المصنفات المستخدمة الّتي عادةً ما تعطي نتائج مميزة في مثل هذا النوع من المهام، لكنّ الطريقة الوحيدة لتحديد المصنف الأمثل لمهمة، ما هي التجربة. نتج، بعد عملية المعالجة المسبقة للبيانات 133 ألف عينة، تم

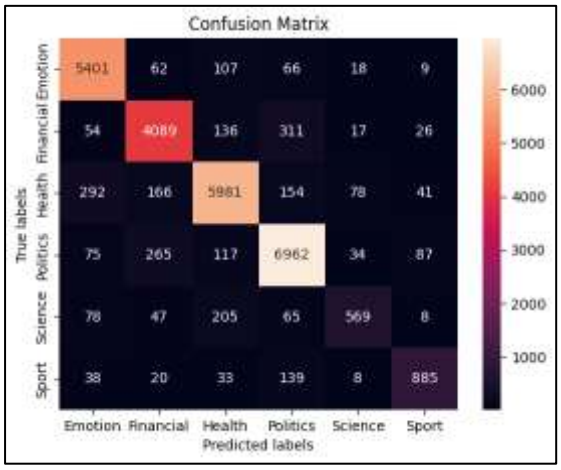
تقسيمها الى مجموعة تدريب Training Set ومجموعة اختبار Test Set بنسبة 80 % الى 20% فحصلنا على 106 ألف عينة لتدريب النماذج وقمنا بتقييم أدائها باستخدام 26 ألف عينة.

توضح النتائج المبينة في الجدول (1) أن نموذج fine-tuned BERTBASE يقدم أداءً أفضل من النماذج الأخرى في كل المقاييس تقريبًا. يمكننا أيضًا أن نلاحظ أن MultinomialNB (alpha=.1) يعطي أعلى قيمة للدقة (Precision). وعلى اعتبار أن مجموعة البيانات غير متوازنة فإن معيار F1−Score هو المعيار الأساسي في تقييم النتائج [11].

الجدول (1): مقارنة بارامترات تقييم الأداء لكافة النماذج التي تم تدريبها

Feature Extraction	Model	Accuracy	Precision	Recall	F1-Score
TF-IDF	LinearSVC (l2 penalty)	0.896	0.891	0.825	0.849
	LinearSVC (l1 penalty)	0.890	0.879	0.820	0.842
	SGDClassifier (l2 penalty)	0.887	0.862	0.821	0.837
	SGDClassifier (l1 penalty)	0.886	0.859	0.818	0.834
	SGDClassifier (Elastic-Net penalty)	0.887	0.862	0.823	0.839
	LogisticRegression	0.868	0.881	0.755	0.790
	MultinomialNB (alpha=.1)	0.850	**0.904**	0.675	0.708
	MultinomialNB (alpha=.01)	0.871	0.887	0.779	0.813
	BernoulliNB (alpha=.1)	0.769	0.836	0.607	0.625
	BernoulliNB (alpha=.01)	0.802	0.780	0.762	0.749
	ComplementNB (alpha=.1)	0.878	0.885	0.792	0.821
	ComplementNB (alpha=.01)	0.871	0.861	0.787	0.813
Word2vec	Bi-directional LSTM with attention	0.877	0.857	0.787	0.812
	$BERT_{BASE}$	**0.897**	0.871	**0.837**	**0.851**

ومن خلال تحليل تقرير التصنيف ومصفوفة الارتباك لنموذج BERT في الشكل (7) والشكل (8)، نلاحظ أن النموذج يمزج بين صنفي "Health" و"Science" بسبب التشابه الكبير في المصطلحات والعبارات المستخدمة في كلا الموضوعين. فعلى سبيل المثال في حال كان المنشور يصف نظام غذائي معين سيتم استخدام مفردات مثل (مقدار البروتين) و (النشاط البدني) وغيرها والتي تعتبر مشتركة بين الموضوعين ما يجعل المصنف متحيزاً للصنف الأكثر تمثيلاً في مجموعة بيانات التدريب. تظهر مصفوفة الارتباك أيضاً تأثير عدم توازن مجموعة البيانات لدينا فالدقة مرتفعة بالنسبة للأصناف ذات التمثيل الأعلى.



الشكل (7): مصفوفة الارتباك لنموذج Fine-tunned BERTBASE

	precision	recall	f1-score	support
0	0.91	0.95	0.93	5663
1	0.88	0.88	0.88	4633
2	0.91	0.89	0.90	6712
3	0.90	0.92	0.91	7540
4	0.79	0.59	0.67	972
5	0.84	0.79	0.81	1123
accuracy			0.90	26643
macro avg	0.87	0.84	0.85	26643
weighted avg	0.90	0.90	0.90	26643

الشكل (8): تقرير التصنيف لنموذج Fine-tunned BERTBASE

2- نتائج سيناريو التصيد الاحتيالي:

تضمنت التجربة عشرة أهداف. سبعة من الأهداف نقرت على زر الاشتراك. لذلك، يمكننا القول إن التصيد الاحتيالي باستخدام الرسائل المخصصة حسب مجال اهتمام الضحية تصل نسبة نجاحه إلى 70%.

الاستنتاجات والتوصيات:

يمكننا أن نستنتج أن استخدام تقنيات الذكاء الاصطناعي في الهندسة الاجتماعية، وخاصة التعلم العميق والمتحولات لتحليل النصوص وفهمها بشكل سياقي يؤدي دورًا حيويًا في اكتشاف اهتمامات مستخدمي وسائل التواصل الاجتماعي ما يجعل مستخدمي هذه المواقع عرضة لهجمات مجرمي الإنترنت ولا سيما هجمات التصيد الاحتيالي.

ونتيجة لذلك، من الضروري تطوير تقنيات جديدة وعالية الجودة لتصفية الرسائل التصيدية بشكل أكثر كفاءة ودقة. ومن ناحية أخرى، هناك حاجة إلى زيادة الوعي المجتمعي حول هذه المشكلة في ظل الاستخدام المتزايد لوسائل التواصل الاجتماعي.

References:

[1] Social Media Stats Syrian Arab Republic | StatCounter Global Stats [Internet]. StatCounter Global Stats. [cited 2023 Nov 5] Available from: https://gs.statcounter.com/social-media-stats/all/syrian-arab-republic/#monthly-202001-202401
[2] National Agency for Network Services. Social Media Cybersecurity [Internet]. [cited 2023 Nov 15] Available from: https://nans.gov.sy/ar/article/social_media_cybersecurity
[3] Alabdan R. Phishing Attacks Survey: Types, vectors, and technical Approaches. Future Internet. 2020 Sep 30;12(10):168. Available from: https://doi.org/10.3390/fi12100168
[4] Phishing Activity Trends Report, 4th Quarter 2022. https://apwg.org/trendsreports/. Anti-Phishing Working Group (APWG); 2022 May.
[5] Alkhalil Z, Hewage C, Nawaf L, Khan I. Phishing Attacks: a recent comprehensive study and a new anatomy. Frontiers in Computer Science [Internet]. 2021 Mar 9;3. Available from: https://doi.org/10.3389/fcomp.2021.563060
[6] Topic_classification_dataset [Internet]. Kaggle. 2023. [cited 2023 Oct 21]. Available from:https://www.kaggle.com/datasets/baraamelhem/topic-classification-dataset/versions/3
[7] Gasparetto A, Marcuzzo M, Zangari A, Albarelli A. A survey on Text Classification Algorithms: From Text to Predictions. Information [Internet]. 2022 Feb 11;13(2):83. Available from: https://doi.org/10.3390/info13020083
[8] Kulkarni A, Shivananda A. Natural Language Processing Recipes: Unlocking Text Data with Machine Learning and Deep Learning using Python [Internet]. 2019. Available from: https://openlibrary.telkomuniversity.ac.id/home/catalog/id/160501/slug/natural-language-processing-recipes-unlocking-text-data-with-machine-learning-and-deep-learning-using-python.html
[9] Grabify. Features - Grabify IP Logger & URL Shortener [Internet]. Features - Grabify IP Logger & URL Shortener. [cited 2024 Jan 8]. Available from: https://grabify.link/faq/features
[10] Tweepy Documentation — tweepy 4.14.0 documentation [Internet]. [cited 2023 Dec 15]. Available from: https://docs.tweepy.org/en/stable/index.html
[11] Bekkar M, Djemaa HK, Alitouche TA. Evaluation Measures for Models Assessment over Imbalanced Data Sets. Journal of Information Engineering and Applications [Internet]. 2013 Jan 1;3(10):27–38. Available from: https://www.iiste.org/Journals/index.php/JIEA/article/download/7633/8051