Journal of Engineering Sciences and Information Technology (JESIT) • Vol 7, Issue 1 (2023) • P: 68 - 58
https://journals.ajsrp.com/index.php/jesit
ISSN: 2522-3321(Online) • ISSN: 2522-3321(Print)

Measure Effectiveness of SMS Spam Detection Model Based on Machine Learning Techniques

Eng. Ahmed Hamed Osman*[1] , Dr. Muhammad Badawi Al-Khalifa[1]

[1] College of Computer Science and Information Technology | Mashreq University | Sudan

Received:
02/01/2023

Revised:
12/01/2023

Accepted:
12/02/2023

Published:
30/03/2023

* Corresponding author:
a.hamid@mashreq.edu.sd

Citation: Osman, A. H., & Al-Khalifa, M. B. (2023). Measure Effectiveness of SMS Spam Detection Model Based on Machine Learning Techniques. *Journal of engineering sciences and information technology, 7(1),* 58 – 68. https://doi.org/10.26389/AJSRP.N020123

Abstract: With the increase in the use of mobile phones, the use of Short Message Service has increased exponentially. With the cost of text messages dropping, people started using them for promotional purposes and unethical activities. This led to a massive increase in spam and consequently the loss of personal and financial data. To prevent data loss, it is essential that spam is detected as quickly as possible. Thus, this paper aims to classify spam not only effectively but also in a short time using python. A dataset containing thousands of text messages containing natural messages (ham) and spam messages was used. Natural language processing techniques were used Multiomail Naive Bayes, Decision Tree and Random Forest are used through which we can classify the message type. After applying these algorithms, Random Forest algorithm got the best accuracy 0.99% in 0.15 second.

Keywords: Accuracy, Classification, Confusion Matrix, Dataset, ham, Natural Language Processing.

قياس فعالية نموذج اكتشاف الرسائل غير المرغوب فيها في خدمة الرسائل القصيرة استناداً على تقنيات التعلم الآلي

م. احمد حامد عثمان*[1]، الدكتور / محمد بدوي الخليفة[1]

[1] كلية علوم الحاسوب وتقانة المعلومات | جامعة المشرق | السودان

المستخلص: مع زيادة استخدام الهواتف المحمولة، زاد استخدام خدمة الرسائل القصيرة بشكل هائل أدى إلى انخفاض تكلفة الرسائل النصية، بدأ الناس في استخدامها لأغراض ترويجية وأنشطة غير أخلاقية. مما أدى ذلك أيضاً إلى زيادة هائلة في الرسائل العشوائية (Spam) وبالتالي يحصل فقدان البيانات الشخصية والمالية. ولمنع فقدان البيانات من الضروري اكتشاف الرسائل العشوائية في أسرع وقت ممكن. تهدف هذه الورقة إلى تصنيف الرسائل العشوائية ليس فقط بشكل فعال، ولكن أيضًا في وقت قصير ، كما انه يعد هذا البحث قابل للتطبيق في الدول الناطقة باللغة الإنجليزية او يتم ارسال الرسائل النصية فيها للمستخدمين باللغة الإنجليزية حتى يومنا هذا.

تم استخدام مجموعة بيانات تحتوي على آلاف الرسائل النصية التي تحتوي على رسائل نصية (*Ham*) ورسائل نصية عشوائية (*Spam*). تم استخدام تقنيات معالجة اللغة الطبيعية وخوارزميات تعلم الآلة (مصنف بايز الساذج(*Naive Bayes*) وشجرة القرار(*Decision Tre*) والغابة العشوائية *(Random Forest)*) التي يمكننا من خلالها تصنيف نوع الرسالة. بعد تطبيق هذه الخوارزميات، حصلت خوارزمية *Random Forest* على أفضل دقة 0.99% في 0.15 ثواني.

الكلمات المفتاحية: الدقة، التصنيف، تعلم الآلة، مجموعة البيانات، عشوائية، معالجة اللغة الطبيعية.

1- المقدمة

في الوقت الحاضر، هناك العديد من شركات الاتصالات في جميع انحاء العالم التي توفر شرائح الاتصال، سواء كانت شريحة إلكترونية مدمجة أو عادية. يمتلك معظم مستخدمي الأجهزة الذكية أو غيرهم هذه الشرائح للتواصل، سواء كان ذلك من خلال المكالمات الصوتية أو خدمة الرسائل القصيرة *Short Message Service* (*SMS*).

الرسائل القصيرة، التي يشار إليها عادةً باسم "الرسائل النصية" هي خدمة لإرسال رسائل قصيرة بين الهواتف المحمولة المختلفة (Grosz, 1982)، وهي جزء من التقنيات المفيدة، ولكن إذا تم استخدامها لأغراض غير مصرح بها مثلا كإرسال إعلانات تجارية وجمع المعلومات واستخدامها هنا قد تصبح تقنية ضارة لبعض المستخدمين.

يقوم المتسللون (الهكر) بإنشاء وتوزيع الرسائل القصيرة والعشوائية للأفراد عبر الأجهزة المحمولة في محاولة لسرقة بياناتهم الحساسة. بالنسبة لأولئك اللذين تصلهم هذه الرسائل العشوائية (*Spam*) التي باعتبارها رسائل عشوائية غير مرغوب فيها بالنسبة للمستلم، إذا التزموا بالتعليمات الواردة في هذه الرسالة وأدخلوا معلوماتهم الشخصية على موقع ويب أو تطبيق مزيف، مثل حسابهم المصرفي عبر الإنترنت، فقد يحصل المتسلل على بعض المعلومات وقد يؤدي ذلك إلى فقدان معلوماتهم المهمة.

في الوقت الحاضر، يكون الاحتيال عبر الرسائل القصيرة العشوائية تأثير كبير على العديد من الأشخاص الذين يصدقون محتويات الرسالة واتباع إرشادات المخترق من خلال النقر على أي روابط زائفة فقد يؤدي ذلك إلى فقدان معلومات مهمة. بالإضافة إلى ذلك، فإن الرسائل القصيرة غير المرغوب فيها مقلقة أكثر بكثير من البريد الإلكتروني العشوائي لأنها تساهم في بعض الدول في نفقات المستلم أيضًا. مع التوافر المحدود لبرامج تصفية الرسائل العشوائية على الهاتف المحمول. تختلف تصفية الرسائل العشوائية في رسائل البريد الإلكتروني والنصوص اختلافا كبيرًا في عدد من الوسائل الرئيسية. تعد قواعد البيانات الحقيقية للرسائل القصيرة غير المرغوب فيها نادرة للغاية، على عكس رسائل البريد الإلكتروني، التي يمكنها الوصول إلى مجموعة متنوعة من مجموعات البيانات الكبيرة. بالإضافة إلى ذلك، نظرًا لأن الرسائل النصية أقصر من رسائل البريد الإلكتروني، فهناك عدد أقل من الصفات التي يمكن الاستفادة منها. كما انه في رسائل الايميل يوجد عنوان للإيميل عكس الرسائل النصية لا يوجد عنوان للرسالة مما يجعل عملية معرفة نوع الرسالة في الرسائل النصية معقد أكثر من رسائل الإيميل.

ولذا يمكن حل هذه المشكلة إذا كانت لدينا أداة مدربة على أنواع هذه الرسائل ويمكنها التعرف بشكل موثوق على الرسائل القصيرة غير المرغوب فيها.

لذلك، سيتم تطوير نموذج يمكنه تصنيف نوع الرسائل وإجراء مدخلات المستخدم لإدخال رسالة واكتشاف ما إذا كانت الرسالة رسالة طبيعية (*Ham*) التي تعتبر رسالة عادية خالية من أنماط الاحتيال او السرقة، أو رسالة عشوائية باستخدام خوارزميات التعلم الآلي، وهي مجموعة من طرق بناء النماذج من البيانات تلقائيًا. تُعرف الخوارزميات التي تحول مجموعة البيانات إلى نموذج باسم خوارزميات التعلم الآلي، وهي أساس التعلم الآلي. تعتمد الخوارزمية المناسبة لاستخدامها اما خاضعة للإشراف (*Supervised*) أو غير خاضعة للإشراف (*Unsupervised*) على نوع المشكلة التي تحاول حلها، وموارد الحوسبة ونوع البيانات التي لديك. أيضًا، سيعتمد هذا المشروع على معالجة اللغة الطبيعية *Natural Language Processing* (*NLP*)، وهي مجموعة تستند إلى أسس نظرية من المناهج الحسابية لقراءة وفهم وتمثيل النصوص التي تحدث بشكل طبيعي على مستوى واحد أو أكثر من التحليل اللغوي، بهدف إنتاج معالجة لغوية ما يرغب فيه الإنسان في مجموعة متنوعة من المهام أو التطبيقات.

تم تقسيم هذه الورقة إلى عدة أجزاء، القسم الأول المقدمة وعرض المشكلة والثاني يناقش الدراسات السابقة والأعمال المتعلقة بنفس الموضوع ويلخص نتائجها ويشرح القسم الثالث الطريقة المستخدمة لبناء النموذج،

من الحصول على البيانات حتى تقييم النتائج ويشرح القسم الرابع عملية معالجة البيانات وبناء النماذج وتقييم الخوارزميات. أخيرًا، يعرض القسم الخامس الاستنتاجات والتوصيات والاقتراحات للدراسات المستقبلية.

2- الدراسات ذات الصلة

في الورقة المكتوبة بواسطة (Ora, 2020)تم استخدام تقنية اختيار ميزة نسبة التردد (*Frequency Ratio Feature Selection*) أثناء تنفيذ الخوارزميات مصنف بايز، الانحدار اللوجستي و شجرة القرار *J- 48* حيث تم تنفيذ تقنية التحقق المتقاطع 10 أضعاف (*10-Fold Cross-Validation*). لقد حصلوا على أعلى دقة من *Naive Bayes* بلغت 94% في وقت قصير.

في تجربة أجريت بواسطة (DasGupta et al., 2021) قرروا تصنيف الخوارزمية بناءً على ميزتين مهمتين: طول الرسالة ومصفوفة ناقل العدد (*Count Vectorizer Matrix*) يستخدمون طول الرسالة كمقياس جودة السمة. خلال تحليلهم الاستكشافي، اكتشفوا أن الرسائل التي تحتوي على رسائل احتيالية لها أطوال مختلفة مقارنة بالرسائل الأخرى غير المقلقة، حيث يمكنهم أن يروا بوضوح أن رسائل (*Ham*) يبلغ طولها 55 حرفًا فقط، مقارنة بمتوسط الرسائل النصية الاحتيالية بطول حوالي 176 حرفًا. يستخدمون "اكتساب المعلومات" (*Information Gain*) كمقياس جودة السمة. يمكن أن يؤدي اكتساب المعلومات التعلم على تصنيف النص إلى تقليل عدد السمات بشكل كبير دون أي خسارة. حققت خوارزمية Naïve Bayes أعلى دقة بلغت % 98.44 باستخدام مصفوفة كسب المعلومات، ومن السهل تصنيف رسالة على أنها رسالة عشوائية أو أنها ليست كذلك.

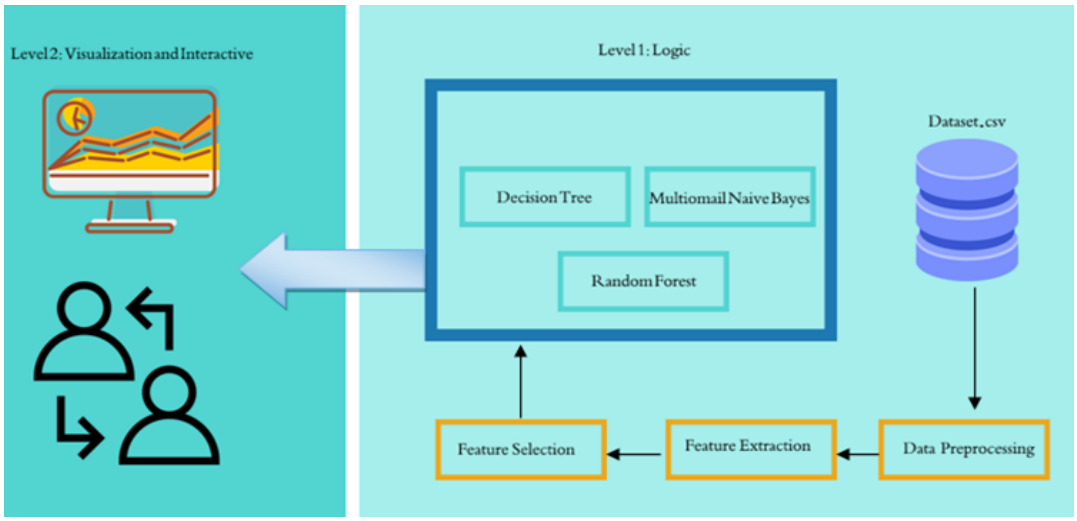
في نموذج الكشف عن منشورات الكراهية باللغة الأمهرية من الفيس بوك (SURAFEL GETACHEW ,2020) ، والتي تعد دراسة فريدة نوعا ما كوناها باللغة الأمهرية التي هي اللغة الرسمية لدولة إثيوبيا وهي لغة الإدارة والمصالح الحكومية كما أنها تستخدم في المدارس الحكومية وخاصة في مراحلها الأولى وتصدر بها غالبية الصحف في إثيوبيا ويستخدمها البعض في مواقع التواصل الاجتماعي. يتم استخدام تمثيل ميزة (*Word2vec Feature Representation*) لتمثيل مجموعة البيانات بواسطة الناقل ولجعلها ملائمة للنموذج عن طريق بناء قاموس يقوم بتعيين الكلمات إلى عدد صحيح. استخدموا شبكة عصبية متكررة تعتمد على الذاكرة طويلة المدى في واحدة من أفضل النتائج، وحصلوا على حوالي 97.9 دقة، وباختبار المستخدم في النموذج وجدوا أن معظم النتائج كانت صحيحة، بناءً على النص إذا كان يحتوي على الكلام الذي يحض على الكراهية كان هذا بعد عدة مرات من التدريب لمجموعة البيانات.

الجدول 1: ملخصات الدراسات ذات الصلة

مساحة للتصنيفات	الخوارزميات المستخدمة	نوع النصوص التي تمت معالجتها	النتائج	الكاتب
تصنيف الرسائل غير المرغوب فيها لخدمة السائل القصيرة	Naive Bayes, J- 48, Logistic Regression	الرسائل القصيرة باللغة الإنجليزية في الهواتف النقالة.	Naive Bayes بدقة وصلت 94%	(Ora, 2020)
تصنيف الرسائل غير المرغوب فيها لخدمة السائل القصيرة	Logistic Regression, Naive Bayes, Random Forest	الرسائل القصيرة باللغة الإنجليزية في الهواتف النقالة.	Naive Bayes بدقة وصلت 98.44%	(DasGupta etal., 2021)
تصنيف خطاب الكراهية	RNN-LSTM.	منشورات باللغة الأمهرية في الفيس بوك	RNN بدقة وصلت 97.9%	(SURAFEL GETACHEW, 2020)

3- المنهجية

تم استخدام المنهج التطبيقي في هذا البحث، حيث انه هو نوع من التصاميم البحثية والذي يسعى إلى حل مشكلة معينة أو تقديم حلول مبتكرة للقضايا التي تؤثر على الفرد أو الجماعات أو المجتمع. غالبًا ما يشار إليها على أنها طريقة علمية للبحث لأنها تتضمن التطبيق العملي للأساليب العلمية على المشكلات اليومية. ويعتمد هذا المشروع على خمس خطوات رئيسية لتصنيف نوع الرسالة والتنبؤ بها، سواء كانت رسالة نصية عادية أو رسالة عشوائية، وهذه الخطوات هي اختيار البيانات، تصحيح البيانات ومعالجتها، استخراج المميزات واختيارها، استخراج البيانات والتقييم.



الشكل 1: منهجية المشروع لاكتشاف الرسائل غير المرغوب فيها عبر الرسائل القصيرة

3.1 اختيار مجموعة البيانات

تحتوي مجموعة البيانات المستخدمة المقدمة من جامعة كاليفورنيا في إيرفين إلى *Kaggle* على 5574 رسالة وتم تنزيل البيانات من موقع *Kaggle* الإلكتروني المتاح بتنسيق *CSV*. مجموعة واحدة من رسائل *SMS* باللغة الإنجليزية تتكون من 5574 رسالة، مقسمة حسب الهام أو البريد العشوائي.

3.2 تنظيف البيانات والمعالجة المسبقة

يتعين علينا تنظيف مجموعة البيانات الخاصة بنا عن طريق إزالة الأحرف غير ذات الصلة وعلامات الترقيم مثل علامة الاستفهام (؟)، والفاصلة (،)، والفاصلة المنقوطة (؛)، والرموز التعبيرية ، والقيم الفارغة ، وعنوان *URL*. بعد ذلك، نطبق تقنيتي المعالجة المسبقة في البرمجة اللغوية التالية:

أ- التطبيع: هو عملية تحويل كلمة أو مصطلح إلى صيغته الأساسية. في عملية التطبيع، تتم إزالة الشكل التصريفي أو الأحرف الزائدة للكلمة، وبالتالي نحصل على الشكل الأساسي للكلمة (Liu et al., 2021).

ب- الترميز: عملية تقسيم كلمة أو جملة إلى أجزاء، تسمى هذه الأجزاء الرموز. على سبيل المثال: "أنا طالب في جامعة المشرق". بعد الترميز ستكون: "انا"، "طالب"، "في"، "جامعة"،"المشرق"، أيضًا يمكن تقسيم هذه الكلمات إلى رمز مميز لكل حرف.

3.3 استخراج المميزات واختيارها

استخراج الميزات: استخلاص المعلومات من مجموعة الميزات الأصلية لإنشاء مساحة فرعية للميزات الجديدة. الفكرة الرئيسية وراء استخراج الميزات هي ضغط البيانات بهدف الحفاظ على معظم المعلومات ذات الصلة

(Ora, 2020)، والتقنية المستخدمة في هذا البحث لاستخراج الميزات هي مصطلح تردد عكس المستند *Term Frequency-Inverse Document Frequency* (*TF-IDF*) إنه مقياس ، يستخدم في مجالات التعلم الآلي ، يمكنه تحديد أهمية أو ملاءمة تمثيلات السلاسل (الكلمات ، العبارات) في مستند بين مجموعة من المستندات، وذلك يمكن استخدامها لتمثيل البيانات النصية في شكل متجه.

اختيار الميزة: هي عملية اختيار أهم الميزات التي سيتم تضمينها في خوارزميات التعلم الآلي وهي واحدة من أهم التقنيات في عمليات التنقيب عن النصوص (Agarwal et al., 2016).

تم استخدام اختبار مربع كاي، وهو اختبار إحصائي يمكننا من خلاله التخلص من الميزة الأقل أهمية عن طريق التحقق من العلاقة بين الميزة التابعة والمستقلة. تعمل Chi-Square بشكل جيد مع مصفوفات TF-IDF لأنها كلها بيانات فئوية.

3.4 تنقيب البيانات

في هذه الخطوة، من أجل الحصول على نتيجة، يتم تنفيذ خوارزميات *Multinomial Naive Bayes* و*Decision Tree* و*Random Forest*. بعد تنفيذ هذه الخوارزميات، سنختار أفضل أداء وننفذ عملية التنبؤ، وهذا يعني أن المستخدم يدخل الرسالة ويتنبأ النموذج بنوع هام الرسالة أو البريد العشوائي.

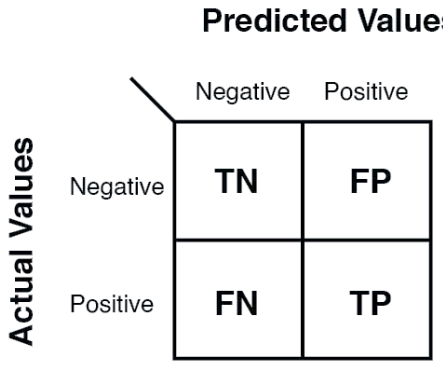
Multinomial Naive Bayes: هو نهج تعليمي احتمالي يستخدم في الغالب في (البرمجة اللغوية العصبية). تعتمد الخوارزمية على نظرية بايز وتحسب احتمالية كل علامة لعينة معينة ثم تعطي العلامة ذات الاحتمال الأعلى كناتج(Heung et al., 2016) .

Decision Tree: هي منهجية تصنيف ، حيث يتم نمذجة عملية التصنيف باستخدام مجموعة من القرارات الهرمية حول متغيرات الميزة ، مرتبة في هيكل مثل الشجرة(Heung et al., 2016).

Random Forest: عبارة عن خوارزمية للتعلم الآلي تم تطويرها بناءً على مجموعة من أشجار القرار. تستخدم هذه الخوارزمية للتصنيف، الانحدار ومهام أخرى. عادةً ما تتمتع خوارزمية الغابة العشوائية بدقة أفضل مقارنة بشجرة القرار. عند استخدام *Decision Tree*، قد تحدث مشكلة فرط في التخصيص ، تجمع خوارزمية *Random Forest* بين *Trees* مختلفة لحل هذه المشكلة (Heung et al., 2016).

3.5 التقييم

بعد تجربة هذه الخوارزميات، يجب تقييم أدائها لمعرفة أفضل أداء. يتم التقييم هنا على عدة معايير تعتمد على مصفوفة الارتباك، وهي أداة لتلخيص أداء خوارزميات التصنيف.



الشكل 2: النموذج العام لمصفوفة الارتباك.

عندما يتم تصنيف المثيلات الإيجابية بشكل صحيح، يُطلق عليها اسم إيجابي حقيقي (*TP*) ، موجب خطأ (*FP*) ، عندما يتم تصنيف المثيلات الإيجابية بشكل غير صحيح ، *False Negative (FN)* ، عندما يتم تصنيف المثيلات

السلبية بشكل غير صحيح و *True Negative (TN)* ، عندما تكون المثيلات السلبية هي مصنفة بشكل صحيح(Choudhary & Jain, 2017).

الدقة *Accuracy*: وهي تحسب الحالات المصنفة الصحيحة من إجمالي الحالات المصنفة.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \qquad (1)$$

الدقة *Precision*: إنها تحسب الحالات التي تم التنبؤ بها بشكل صحيح والتي تبين بالفعل أنها إيجابية.

$$Precision = \frac{TP}{TP + FP} \qquad (2)$$

الاستدعاء *Recall*: يحسب الحالات الإيجابية الفعلية التي تمكنا من التنبؤ بها بشكل صحيح باستخدام نموذجنا.

$$Recall = \frac{TP}{TP + FN} \qquad (3)$$

المقدار *F1-Score*: هو قياس الدقة الكلية للخوارزمية التي تجمع بين الدقة والاسترجاع.

$$F1 = 2 * \frac{Precision * Recall}{Precision * Recall} \qquad (4)$$

4- النموذج والنتائج المقترحة

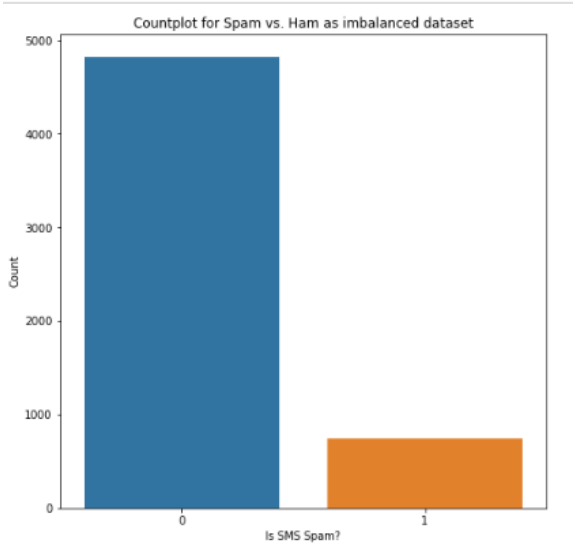
تعتمد معظم العمليات في لغة بايثون على استخدام المكتبات التي يتم استدعاؤها والعمل عليها برمجيًا، انظر الجدول أدناه للحصول على تفاصيل المكتبات المستخدمة.

الجدول 2: مكتبات بايثون المستخدمة.

اسم الاداة	الاصدار	الوصف	الموقع الرسمي
Numpy	1.23.1	معالجة المصفوفة للأرقام والسلاسل والكائنات. نستخدمها للتعامل مع ميزات مجموعة البيانات الخاصة بنا للتدريب واختبار النموذج.	/https://numpy.org
Matplotlib	3.5.1	احد أهم لمكتبات لعرض البيانات على شكل صور	https://matplotlib.org/
NLTK	3.6.5	برنامج بايثون يساعد على العمل مع بيانات اللغة الطبيعية. نحن نستخدمه في مهام المعالجة المسبقة للبيانات.	/https://www.nltk.org
Pandas	1.4.3	أداء عالي وسهل الاستخدام لهياكل وأدوات لتحليل البيانات. نستخدمها لقراءة البيانات ومعالجتها وكتابتها والتعامل معها.	/https://pandas.pydata.org
Seaborn	0.11.2	مكتبة تستخدم Matplotlib في الأسفل لرسم الرسوم البيانية. سيتم استخدامه لتصور التوزيعات العشوائية	/https://seaborn.pydata.org

4.1 تحليل البيانات الاستكشافية:

يتضمن تحليل البيانات الاستكشافية استخدام الرسومات والتصورات لاستكشاف مجموعة البيانات وتحليلها. إنها أفضل ممارسة لفهم البيانات ثم تنفيذ عملية التنقيب عن البيانات. تم استخدام مكتبة Matplotlib مع Word Cloud لعرض الكلمات الأكثر شيوعًا في رسائل البريد العشوائي والرسائل الطبيعية في مجموعة البيانات.



الشكل 3: نوع الرسائل في مجموعة البيانات

4.2 أخذ عينات البيانات:

مشكلة التصنيف غير المتوازن هي ما نواجه به عندما يكون هناك انحراف شديد في توزيع بيانات التدريب الخاصة بنا. وهو نهج لمكافحة هذا التحدي هو أخذ العينات العشوائية(Baker, 2017) . هناك طريقتان رئيسيتان لإجراء إعادة التشكيل العشوائي:

الإفراط في أخذ العينات: -نسخ عينات من الفئة الأقلية.

نقص أخذ العينات: -حذف عينات من فئة الأغلبية(Ouda, 2017). في هذه الورقة، يتم إجراء عملية أخذ عينات عشوائية باستخدام أخذ العينات الزائدة

4.3 تدريب واختبار مجموعة البيانات

في التعلم الآلي واستخراج البيانات، يتم تقسيم مجموعات البيانات إلى مجموعتين فرعيتين. تُعرف المجموعة الفرعية الأولى باسم بيانات التدريب -وهي جزء من مجموعة البيانات الفعلية التي يتم إدخالها في نموذج التعلم الآلي لاكتشاف الأنماط وتدريب النموذج عليها.

تُعرف المجموعة الفرعية الأخرى باسم بيانات الاختبار. بعد إنشاء نموذج التعلم الآلي، تحتاج إلى بيانات التي لم يتم استخدامها في عملية التدريب لاختبار النموذج. تسمى هذه البيانات بيانات الاختبار (Ma et al., 2016)، ويمكنك استخدامها لتقييم أداء الخوارزميات التي قمنا باستخدامها. هنا في هذا النموذج تم تقسيم مجموعة البيانات على النحو التالي: 80% للتدريب و20% للاختبار.

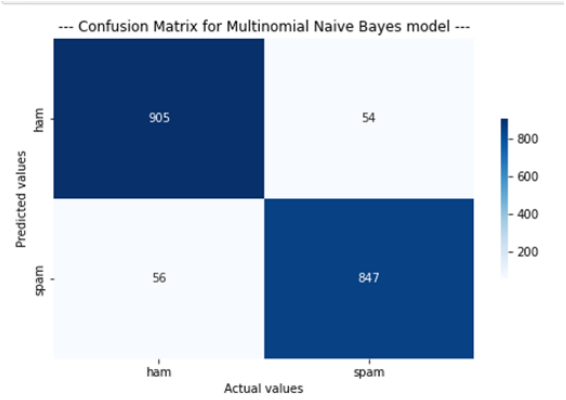
4.4 بناء النموذج والتقييم

كما ذُكر سابقًا، سيتم اختيار النموذج بعد تجربة عدة خوارزميات واختيار أفضل نتيجة بينها، ومن ثم إجراء التنبؤ في الخطوة الأخيرة.

4.4.1 التجربة 1: نتائج النموذج باستخدام خوارزمية *Multinomial Naïve Bayes*

الجدول 3: نتاج التجربة 1.

F1-Score	Recall	Precision	Accuracy
0.94%	0.94%	0.94%	0.94%

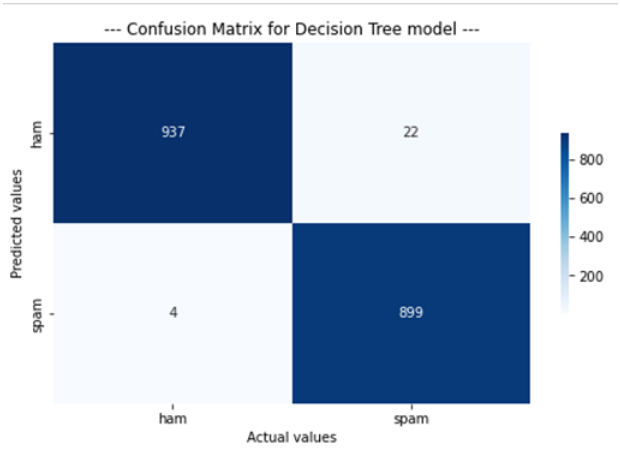


الشكل 4: مصفوفة الارتباك (Multinomial Naïve Bayes)

4.4.2 التجربة 2: نتائج النموذج باستخدام خوارزمية (*Decision Tree*)

الجدول:4 نتائج التجربة 2.

F1-Score	Recall	Precision	Accuracy
0.99%	0.99%	0.97%	0.98%

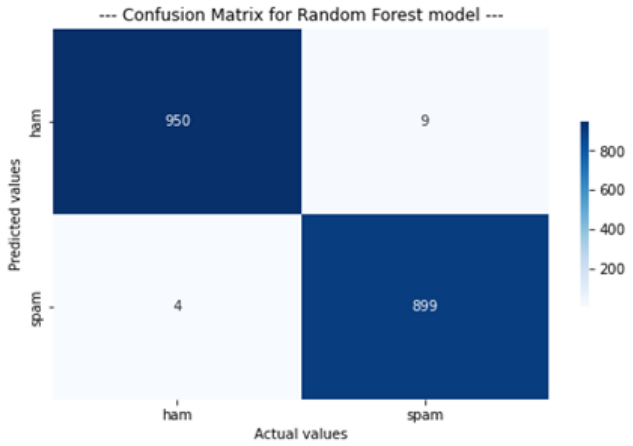


الشكل 5: مصفوفة الارتباك (*Decision Tree Model*)

4.4.3 التجربة 3: نتائج النموذج باستخدام خوارزمية (Random Forest)

جدول 5: نتائج التجربة 3.

F1-Score	Recall	Precision	Accuracy
0.99%	0.99	0.99%	0.99%



الشكل:6 مصفوفة الارتباك (Random Forest)

جدول 6: ملخص تجربة الخوارزميات.

الخوارزمية	الدقة Accuracy	الدقة Precision	الاستدعاء Recall	المقدار F1-Score
Multinomial Naïve Bayes	0.94%	0.94%	0.94%	0.94%
Decision Tree	0.98%	0.97%	0.99%	0.99%
Random Forest	0.99%	0.99%	0.99%	0.99%

بعد تجربة الخوارزميات الثلاثة السابقة، حصلت خوارزمية *Random Forest* على أعلى دقة قدرها 0.99%وإجمالي عدد 13 رسالة تم تصنيفها بشكل خاطئ، وهذه أقل قيمة مقارنة بالخوارزميات الأخرى في هذا البحث، بناءً على هذه النتائج سيتم اجراء عملية بالتنبؤ.

4.5 عملية التنبؤ

في (الشكل 6)، قمنا ببناء دالة برمجية يمكن للمستخدم إدخال الرسالة ويتنبأ النموذج بنوع تلك الرسالة. كما يوجد ثلاث تجارب وجميع المخرجات كانت صحيحة بناء على البيانات المستخدمة.

الشكل 7: الشفرة البرمجية لعملية التنبؤ.

الجدول 7: مقارنة النماذج السابقة مع النموذج الحالي.

الكاتب	النتائج	الخوارزميات المستخدمة	مساحة للتصنيفات
(Ora, 2020)	Naive Bayes بدقة وصلت 94%	Naive Bayes, J- 48, Logistic Regression	تصنيف الرسائل غير المرغوب فيها لخدمة السائل القصيرة
(DasGupta etal., 2021))	Naive Bayes بدقة وصلت 98.44%	Logistic Regression, Naive Bayes, Random Forest	تصنيف الرسائل غير المرغوب فيها لخدمة السائل القصيرة
(SURAFEL GETACHEW, 2020)	RNN بدقة وصلت 97.9%	RNN-LSTM.	تصنيف خطاب الكراهية
الدراسة الحالية	Random Forestبدقة وصلت 99.3%	Multinomial Naive Bayes, Decision Tree, Random Forest	تصنيف الرسائل غير المرغوب فيها لخدمة السائل القصيرة

5. الخلاصة والتوصيات والاقتراحات للدراسات المستقبلية

يمثل البريد العشوائي مشكلة خطيرة، وهي تزداد خطورة كل يوم. قد تكون مشكلة أكثر خطورة للأشخاص الذين لا يعرفون أبدًا ما هو البريد العشوائي. قمنا ببناء هذا النموذج وكتبنا هذه الورقة لتصنيف هذه الرسائل في أسرع وقت وأدق نتيجة. تم استخدام مجموعة بيانات تحتوي على رسائل نصية وتم استكشاف هذه الرسائل لمعرفة ما إذا كانت تحتوي على قيم خاطئة أو مفقودة. تمت معالجة البيانات باستخدام مكتبة NLTK وهي مجموعة أدوات اللغة الطبيعية التي تساعدنا في أزله الرموز الخاصة وقمنا بتحويل الأحرف من الأحرف الكبيرة إلى الأحرف الصغيرة لتقليل الأخطاء والضوضاء. تم استخدام *Word Cloud* لعرض الرسائل الأكثر شيوعًا في مجموعة البيانات. قسمنا مجموعة بيانات إلى 20٪ للاختبار و 80٪ للتدريب، ثم استخدمنا الخوارزميات الثلاثة *Multinomial Naive Bayes* و

Decision Tree و *Random Forest*، حصلت *Random Forest* على أعلى دقة تصل إلى 0.99 في 0.15 ثانية، ثم قمنا ببناء دالة برمجية يمكن التنبؤ بأي رسالة من خارج مجموعة البيانات وقمنا باختبار أداء النموذج داخل الوظيفة حيث كانت جميع النتائج صحيحة لنوع الرسالة. للعمل المستقبلي، محاولة تحسين النتائج، يجب استخدام مجموعة بيانات أخرى وتدريب النموذج عليها. تعد معالجة البيانات مهمة للغاية، فعند معالجة البيانات بشكل متكرر ستحقق نتائج أفضل. يقتصر هذا النموذج على مجموعة بيانات باللغة الإنجليزية فقط. لزيادة حدود البحث، يجب استخدام مجموعة بيانات تحتوي على رسائل بلغات أخرى. سيكون من الرائع أن يتم تطوير هذا النموذج ليكون بمثابة تطبيق واجهة مستخدم رسومية.

تتضمن تقنية تمثيل *TF-IDF* فقط معلومات حول المصطلحات والترددات المقابلة لها في مستند مستقل عن مواقعها في الجملة أو المستند. لذلك، يوصى في المستقبل باستخدام طرق تصنيف النص الدلالي.

REFERENCES

- Agarwal, S., Kaur, S., & Garhwal, S. (2016). SMS spam detection for Indian messages. Proceedings on 2015 1st International Conference on Next Generation Computing Technologies, NGCT 2015, September, 634–638. https://doi.org/10.1109/NGCT.2015.7375198
- Baker, S. (2017). Semantic text classification for cancer text mining. September, 171.
- Choudhary, N., & Jain, A. K. (2017). Towards filtering of SMS spam messages using machine learning based technique. Communications in Computer and Information Science, 712, 18–30. https://doi.org/10.1007/978-981-10-5780-9_2
- DasGupta, S., Saha, S., & Das, S. K. (2021). SMS spam detection using machine learning. Journal of Physics: Conference Series, 1797(1). https://doi.org/10.1088/1742-6596/1797/1/012017
- Grosz, B. J. (1982). Natural language processing. Artificial Intelligence, 19(2), 131–136. https://doi.org/10.1016/0004-3702(82)90032-7
- Heung, B., Ho, H. C., Zhang, J., Knudby, A., Bulmer, C. E., & Schmidt, M. G. (2016). An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma, 265, 62–77. https://doi.org/10.1016/j.geoderma.2015.11.014
- Ideo, R. E. D. I. N. ν, & Ystems, S. U. S. (2017). Master ' s Thesis by. November, 1–73.
- Liu, X., Lu, H., & Nayak, A. (2021). A Spam Transformer Model for SMS Spam Detection. IEEE Access, 9, 80253–80263. https://doi.org/10.1109/ACCESS.2021.3081479
- Ma, J., Zhang, Y., Liu, J., Yu, K., & Wang, X. (2016). Intelligent SMS spam filtering using topic model. Proceedings - 2016 International Conference on Intelligent Networking and Collaborative Systems, IEEE INCoS 2016, 380–383. https://doi.org/10.1109/INCoS.2016.47
- Ora, A. (2020). Spam Detection in Short Message Service Using Natural Language Processing and Machine Learning Techniques.Master's Dissertation, National College of Ireland, Dublin, Ireland.
- Ouda, A. M. N. A. (2017). Arabic SMS Spam Detection Based on Semantic Classification.
- Tesayc, S. G.(2020). Developing Automated Amharic Hate Speech Posts Detection Model From Facebook Using Deep Learning. Master's Dissertation, Addis Ababa Science and Technology University, Addis Ababa, Ethiopia .